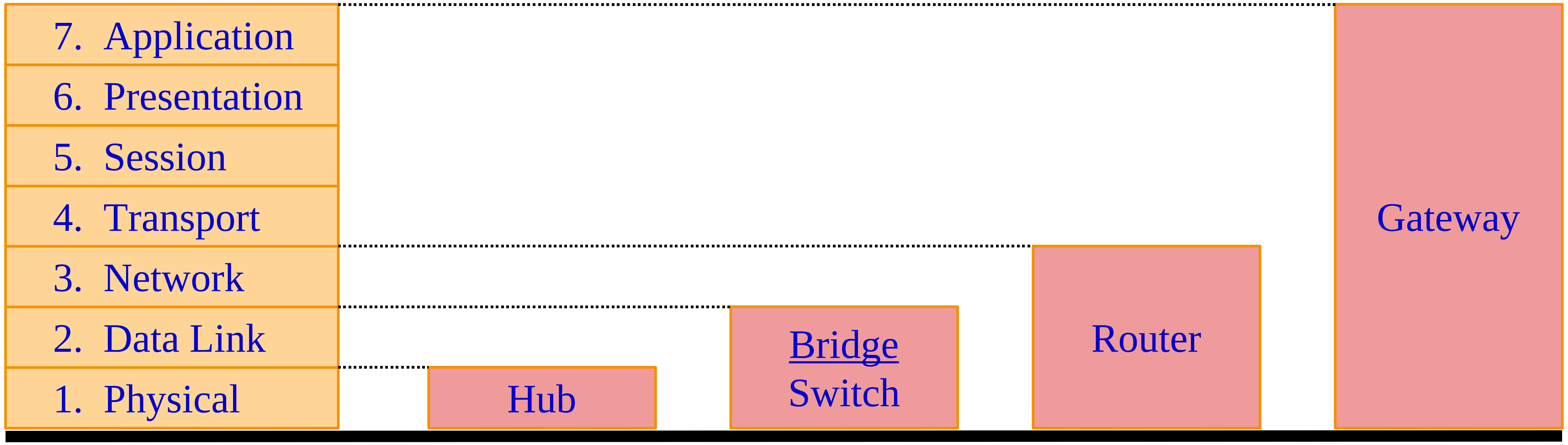
Basic Requirements of a Switch Router

What does / can a router do ?

What can be done in hardware / software ?

How do protocols / standards influence the design ?

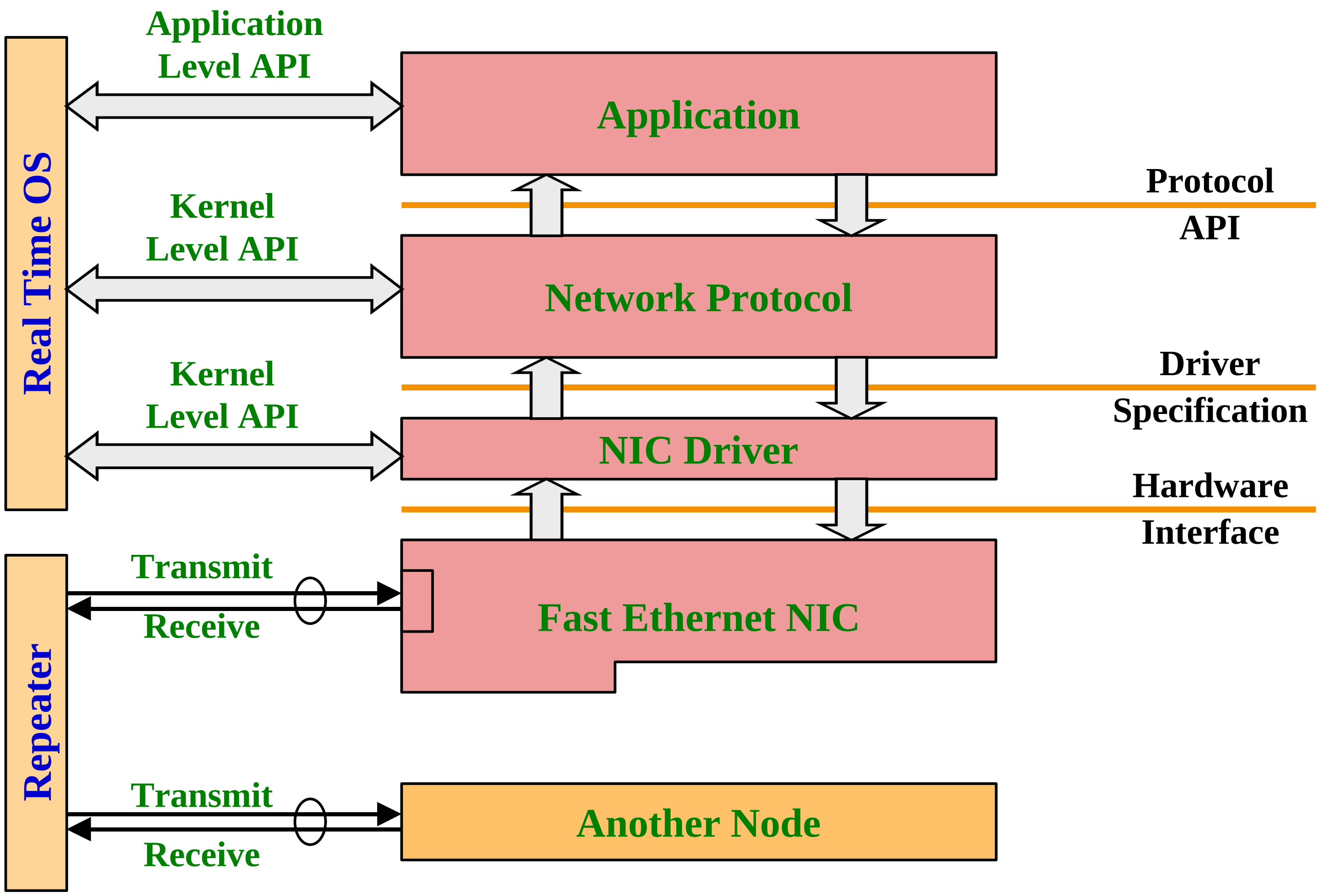
Types of Inter-network Nodes



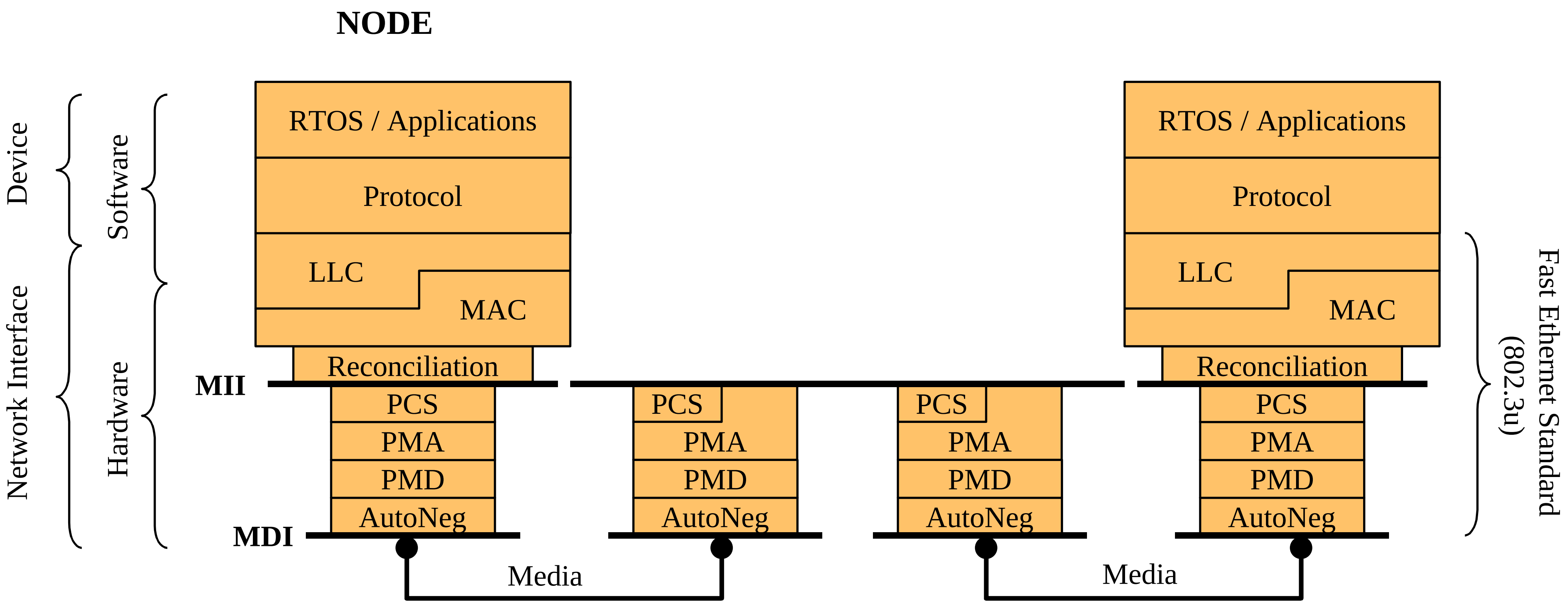
What's the difference? Switch vs Router

Layer	*Purpose of the Layer*	*Role of Switching*
(7) Application (6) Presentation (5) Session	*Defines user-oriented services such as file transfer, messaging, and transaction processing; provides for structuring applications, coding the data, and exchanging information*	*Application switching (e.g. e-mail forwarding); gateways between different application types; support for management functions; selection of destination for messages*
(4) Transport	*Delivery of data to applications, division of messages into packets*	*Directs the messages to the specific destination application or protocol type*
(3) Network	*End-to-end communications through one or more subnets; selects optimal routes; controls loops; manages addressing*	*Forwards packets through an interconnected set of networks*
(2) Data Link	*Transfer of frames across a single network link such as a LAN; manages contention*	*Controls switched circuits, switched LANs, and recovers from link errors*
(1) Physical	*Transmission over a physical circuit including physical connectors, bit encoding, etc.*	*Circuit switching as is used for telephony and port switching for LAN physical media*

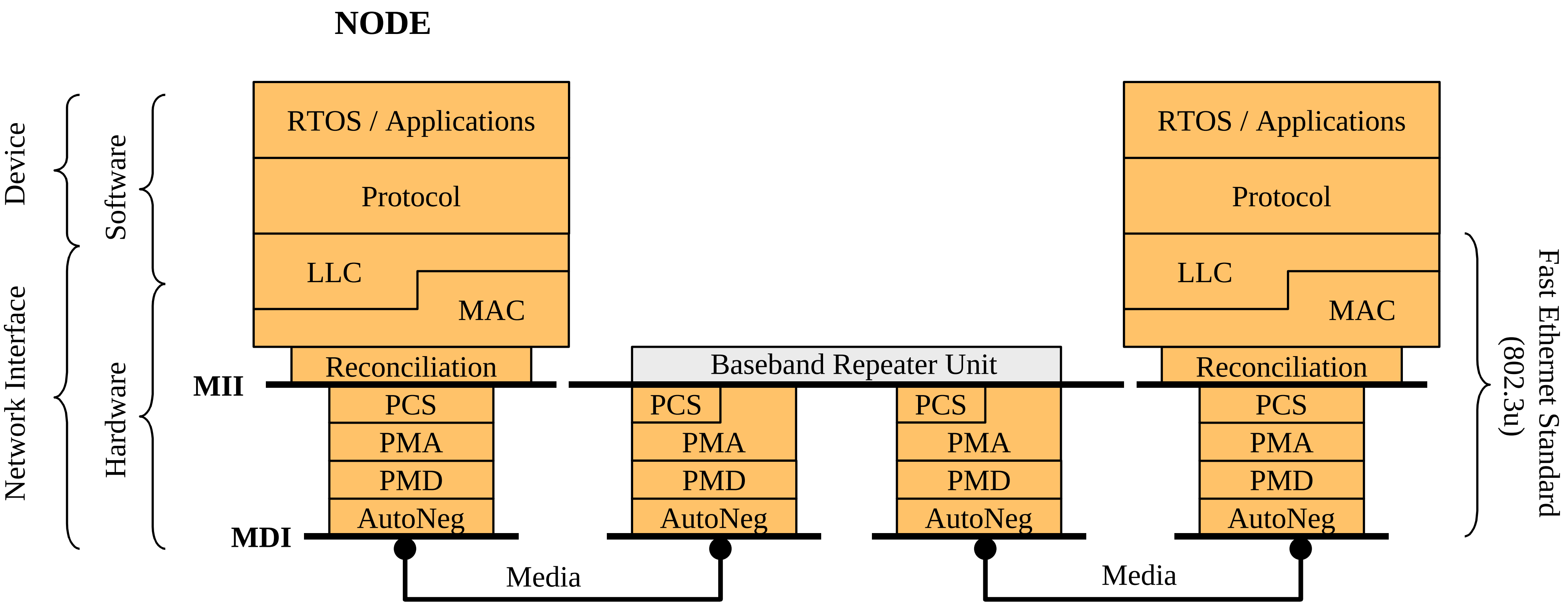
Anatomy of a Node



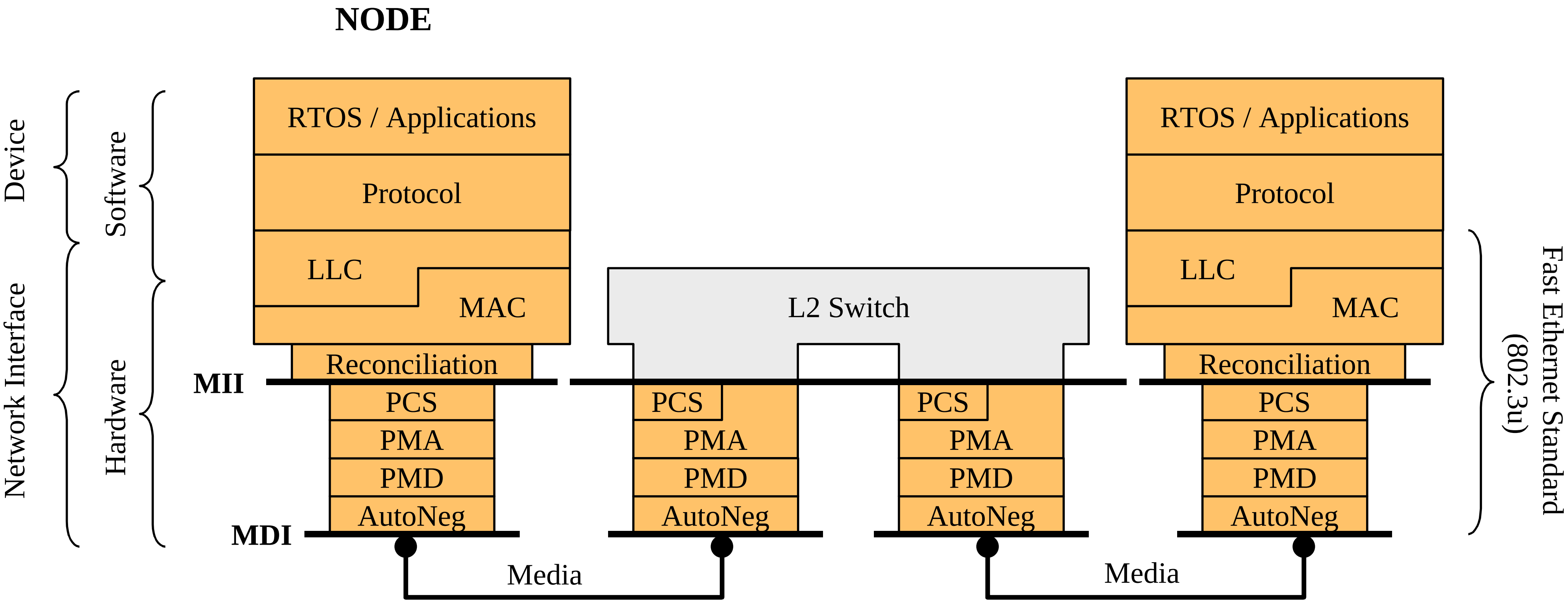
Fast Ethernet System



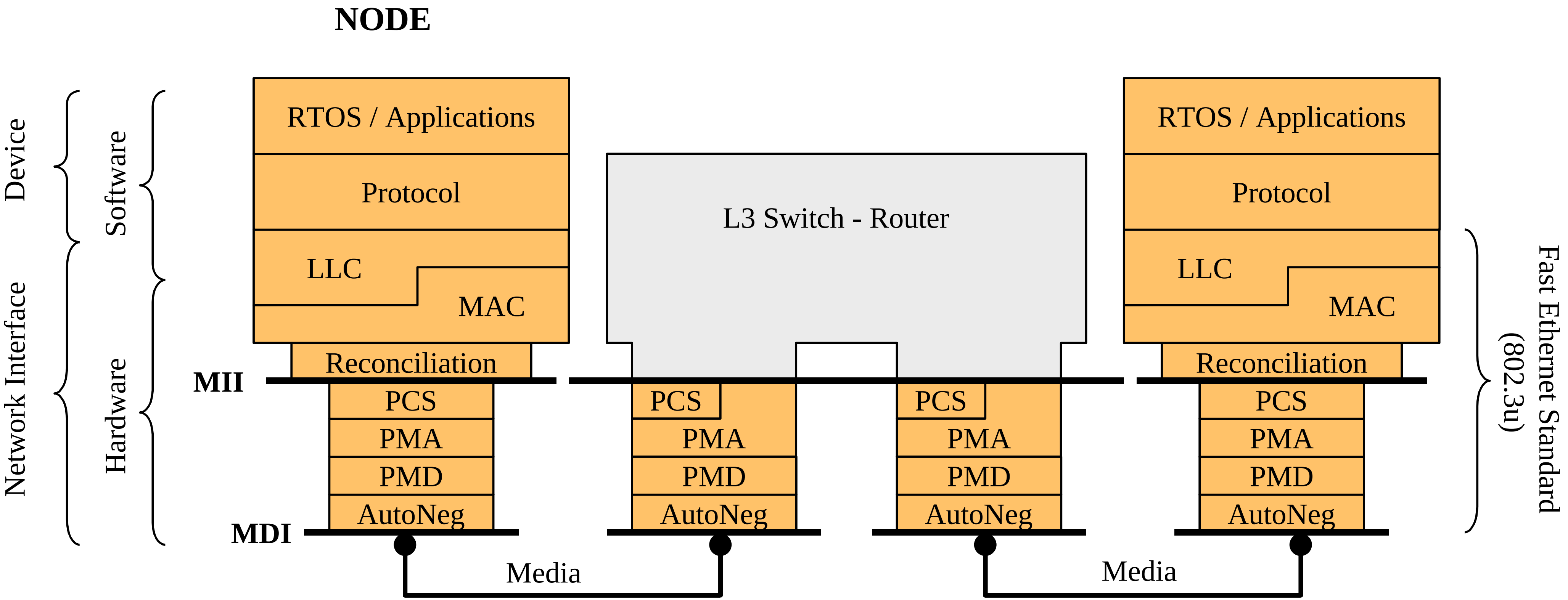
Fast Ethernet System



Fast Ethernet System



Fast Ethernet System



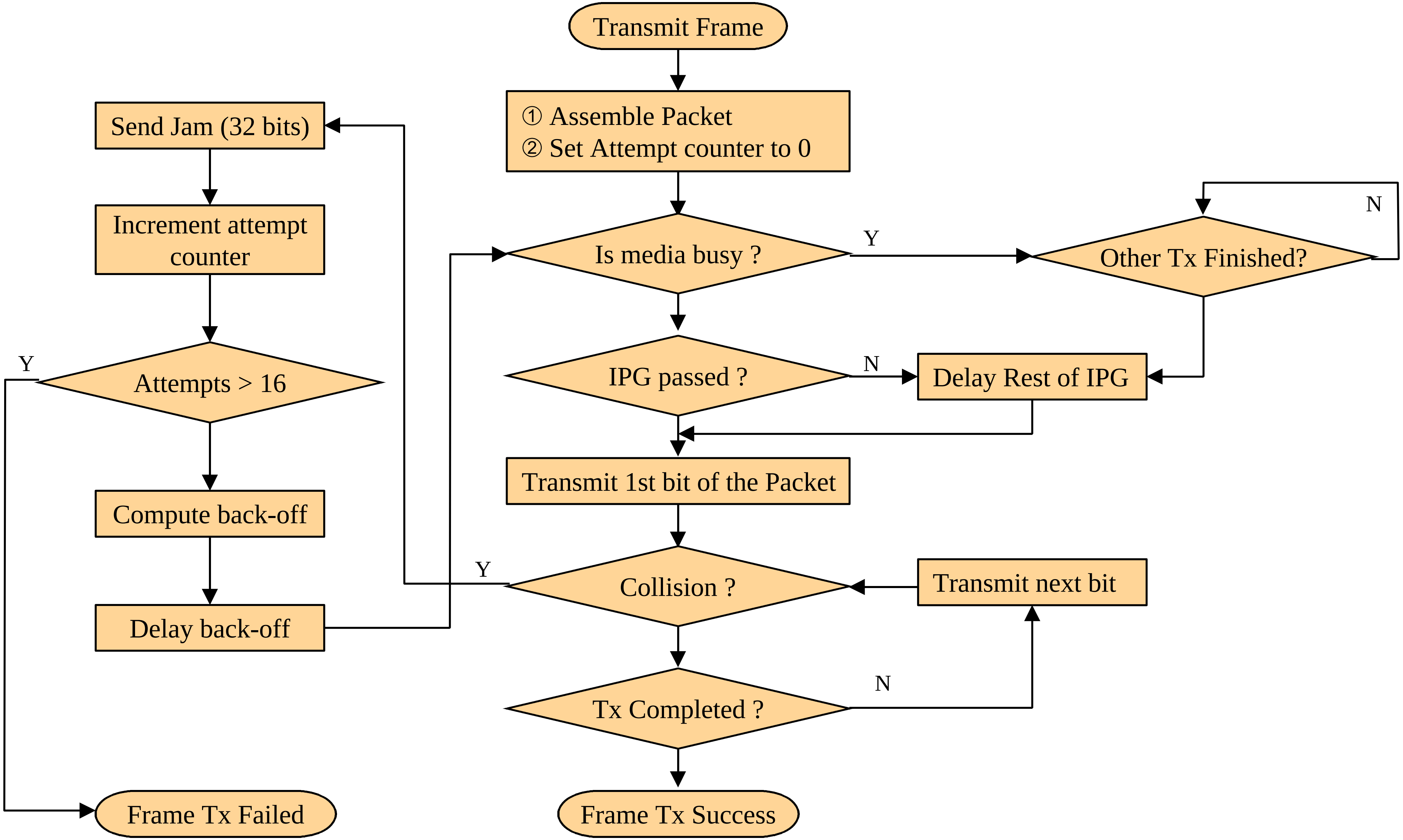
Media Access Controller

- Determine when a node can transmit a packet
- Send frames to the PHY for conversion into packets and transmission on the media
- Receive frames from the PHY and send them to the software that processes frames (protocols and applications).
- Frame checking
  - » Valid Frames
    - – Frame size between 64 bytes & 1518 bytes
    - – Valid frame check sequence (CRC)
    - – Even number of octets
  - » Non-valid Frames
    - – Runts: Any frame that is shorter than 64 bytes (512 bits) in size
    - – Jabber: Data transmission greater than 400 ms (largest packet: 120.56 ms)
    - – Dribble: Invalid number of octets
- Media independent

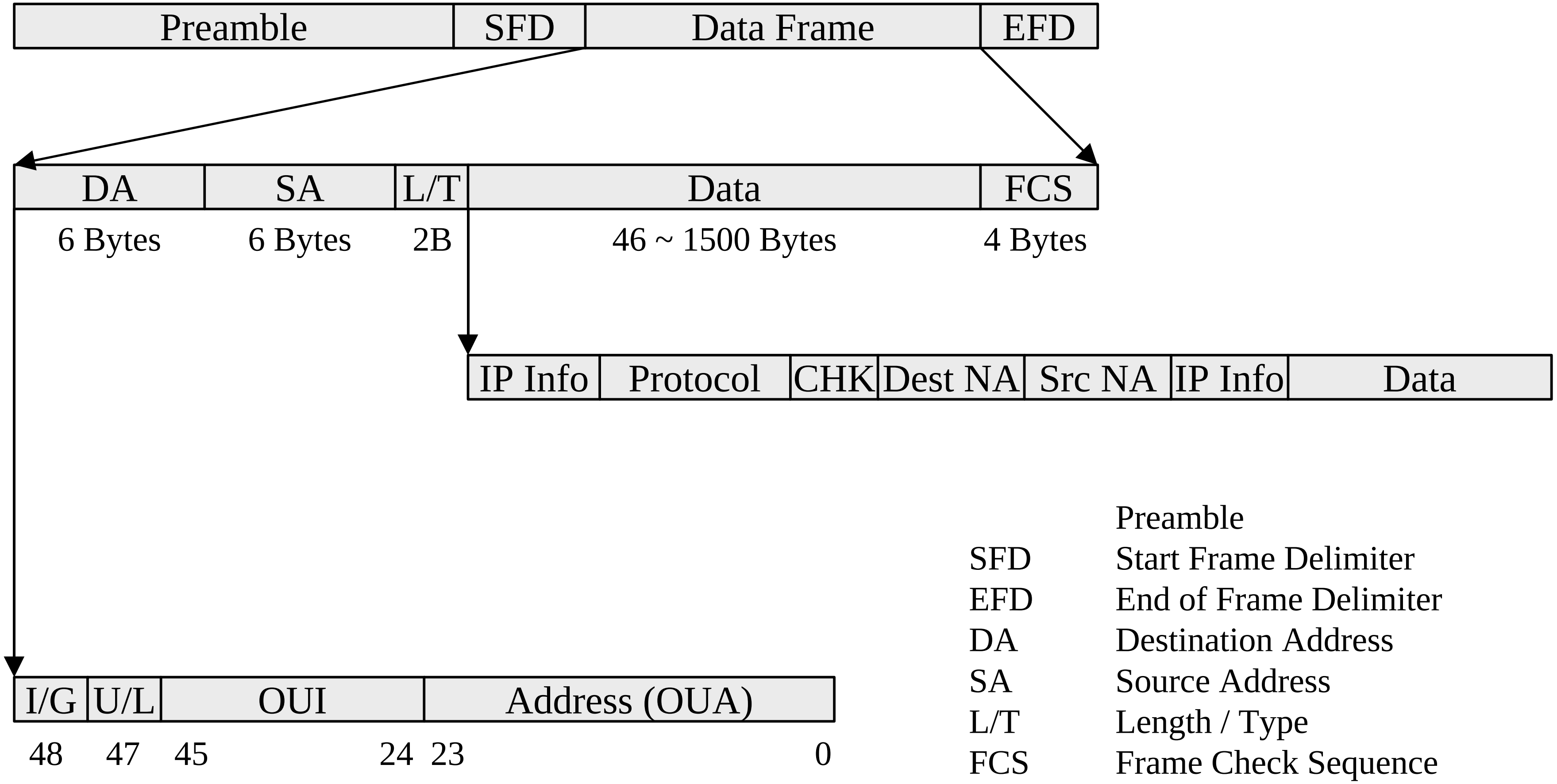
IEEE 802.3 — CSMA / CD

- Media Access Rules
  - Listen before sending
    - CSMA — Carrier Sense Multiple Access
    - Interpacket Gap (IPG = 96 bit time or 0.96 µs for FE)
  - Backoff
    - CD — Collision Detection
    - Collision domain
    - Collision window
    - Slot time == maximum allowable collision window (512 bit times)
      - minimum frame size (512 bit / 64 bytes)
      - maximum network diameter
    - Truncated binary exponential backoff
      - RAND(0, $2^{\min(N,10)}$), where N is the transmit attempt counter
      - Integer multiple of 512 bit slot time (i.e. 512, 1024, 1536, 2048, … , 4096, etc.)
      - Maximum backoff time is 5.3 ms.

CSMA / CD Flow Chart



Packet Format



Preamble | SFD | Data Frame | EFD

DA | SA | L/T | Data | FCS

6 Bytes | 6 Bytes | 2B | 46 ~ 1500 Bytes | 4 Bytes

IP Info | Protocol | CHK | Dest NA | Src NA | IP Info | Data

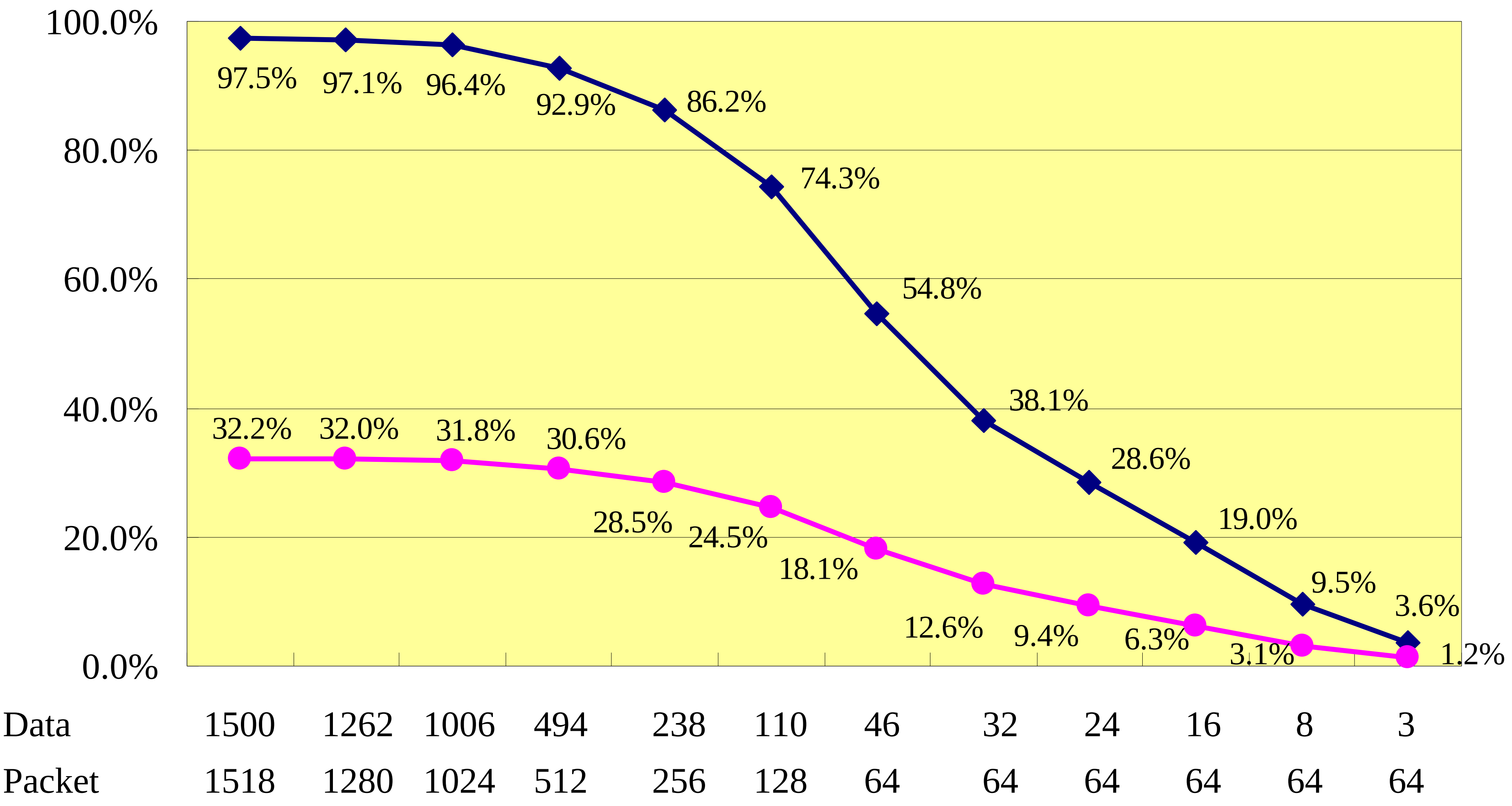
I/G | U/L | OUI | Address (OUA)

48 | 47 | 45 | 24 | 23 | 0

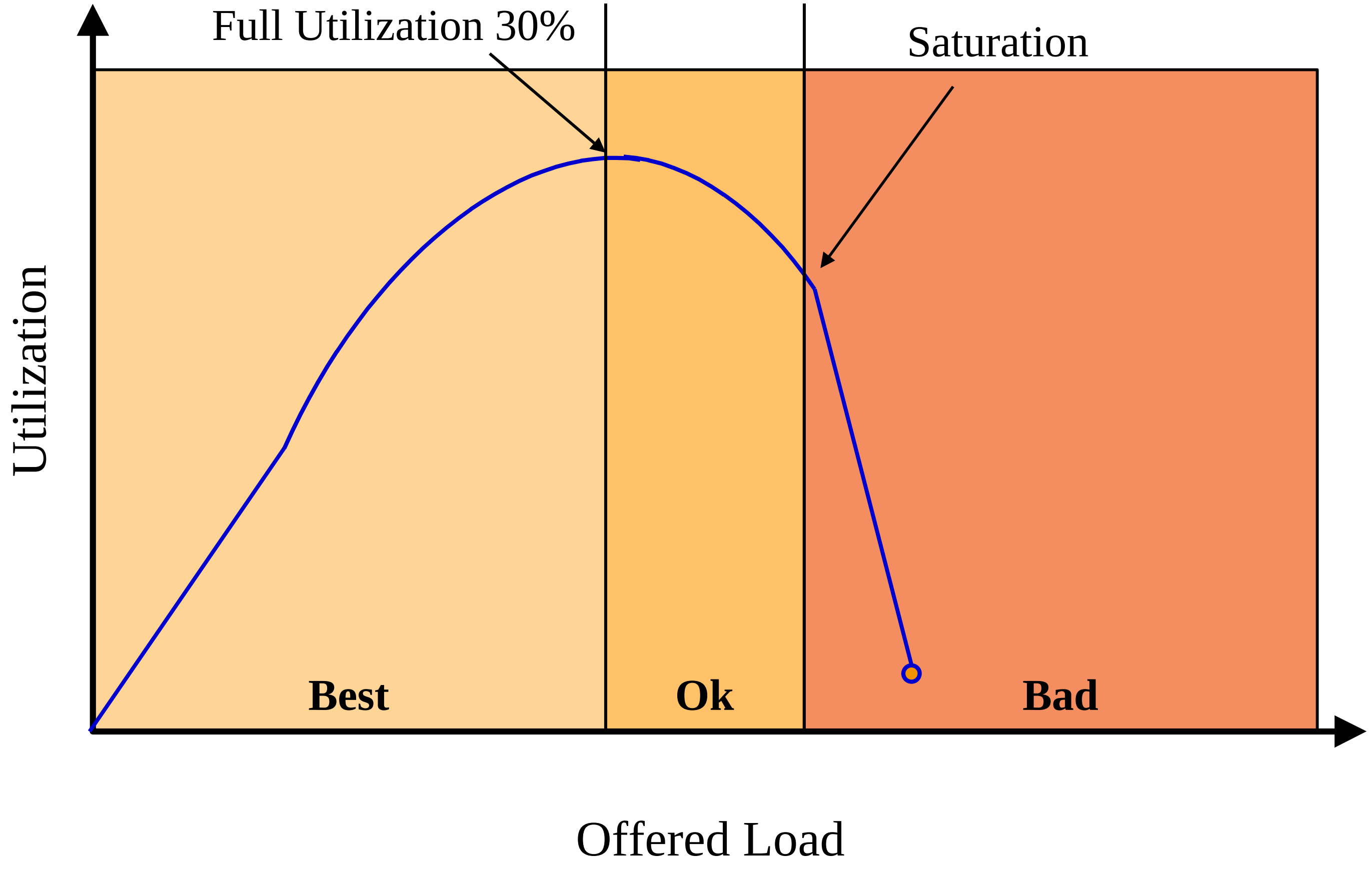
	Preamble	10101010
SFD	Start Frame Delimiter	10101011
EFD	End of Frame Delimiter	----
DA	Destination Address	
SA	Source Address	
L/T	Length / Type	
FCS	Frame Check Sequence	
I/G	Individual / Group	
U/L	Universal / Local Administration	
OUI	Organizationally Unique Identifier	
OUA	Organizationally Unique Address	

ITRI
CCL

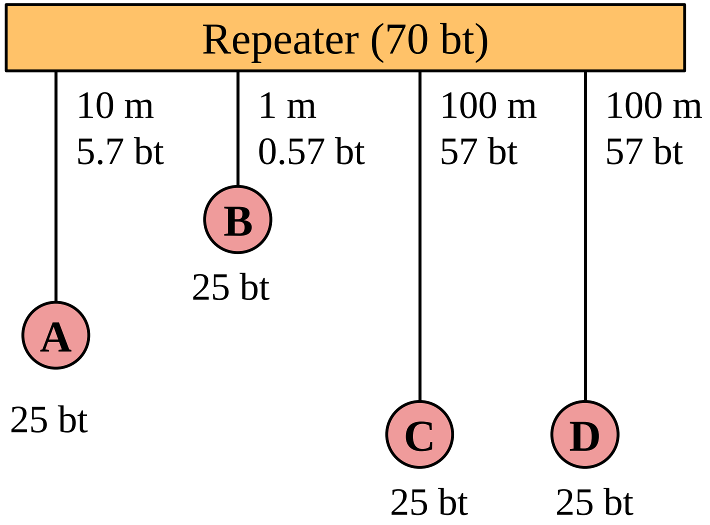
Wire Speed — Efficiency



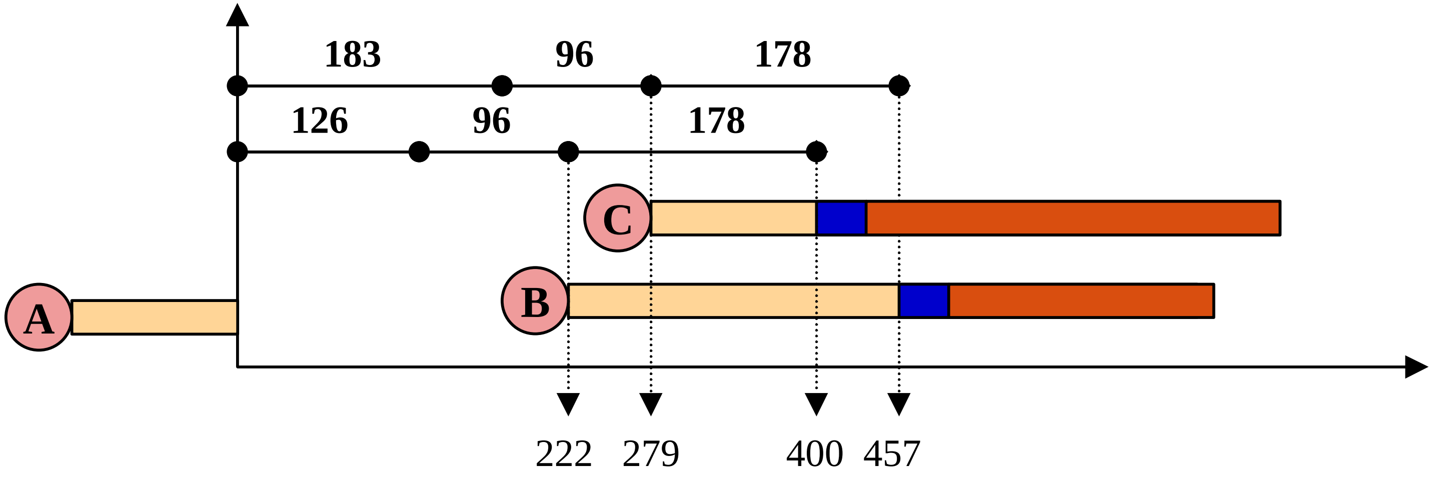
Network Utilization



Basic & Worst Case Collision Detection w/ Cat-5 Cable

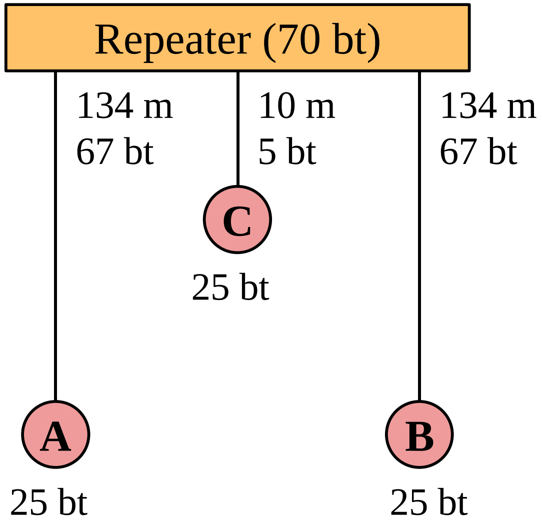


Node-to-node	**Path Delay Value** **(rounded to the nearest whole bit time)**
A → B	126 = 25 + 5.7 + 70 + 0.57 + 25
A → C	183 = 15 + 5.7 + 70 +57 + 25
B → C	178 = 25 + 0.57 + 70 +57 +25
C → D → C	468 = 2 * (25 + 57 + 70 + 57 + 25)
Safety margin	4
Bit-time margin	40 = 512 - 472 -4

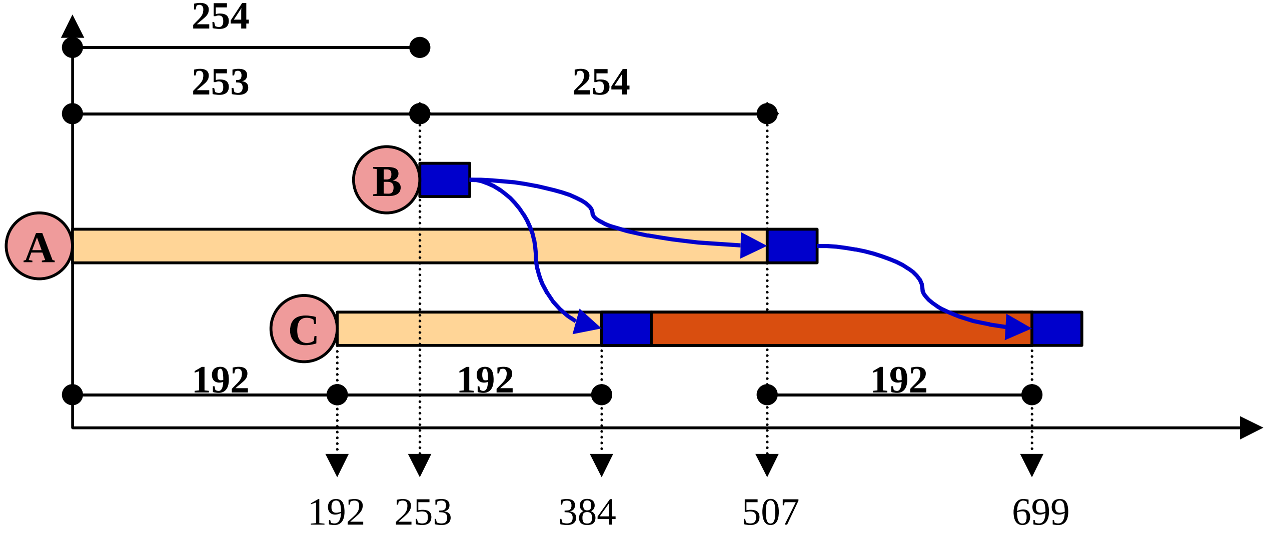


ITRI
CCL

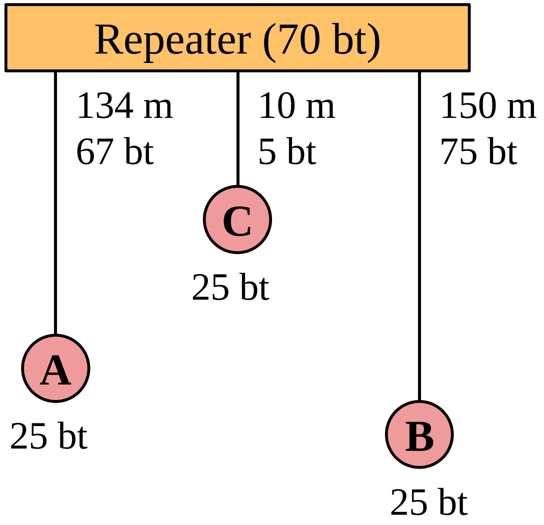
Worst-case Collision Window w/ Class-I Hub and Fiberoptic Cable



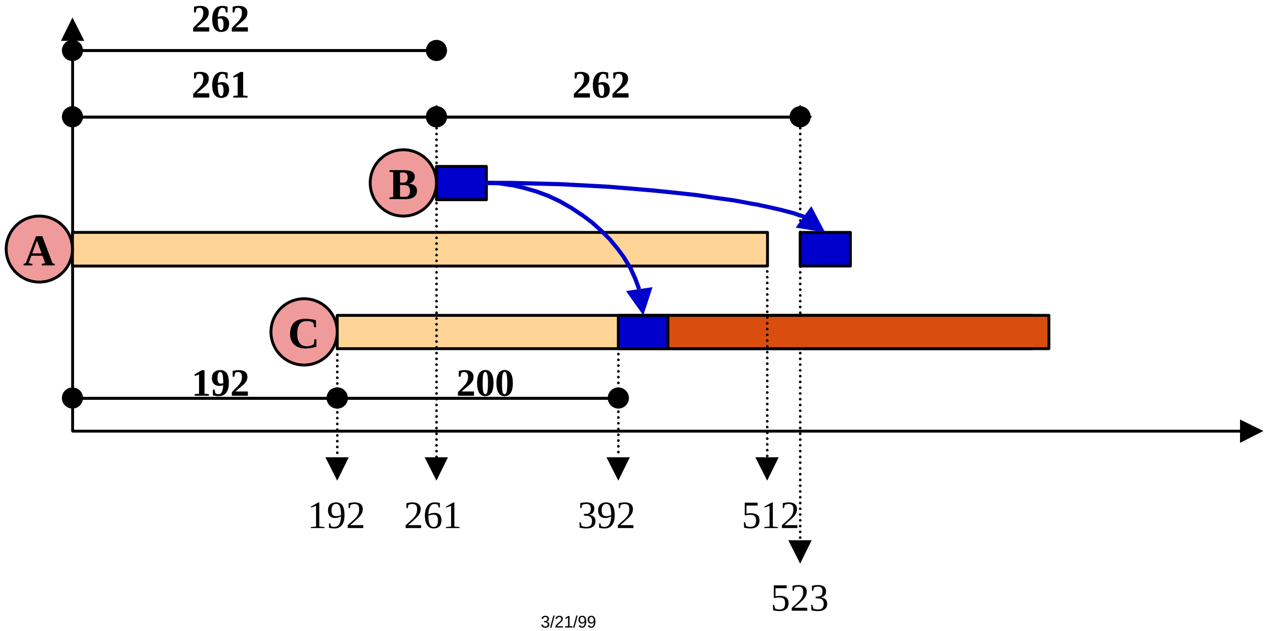
Node-to-node	Path Delay Value (rounded to the nearest whole bit time)
A → B	254 = 25 + 67 + 70 + 67 + 25
A → C	192 = 15 + 67 + 70 + 5 + 25
B → C	192 = 25 + 5 + 70 + 67 +25
A → B → A	508 = 2 * (25 + 67 + 70 + 67 + 25)
Safety Margin	4
Bit time margin	0 = 512 -508 - 4



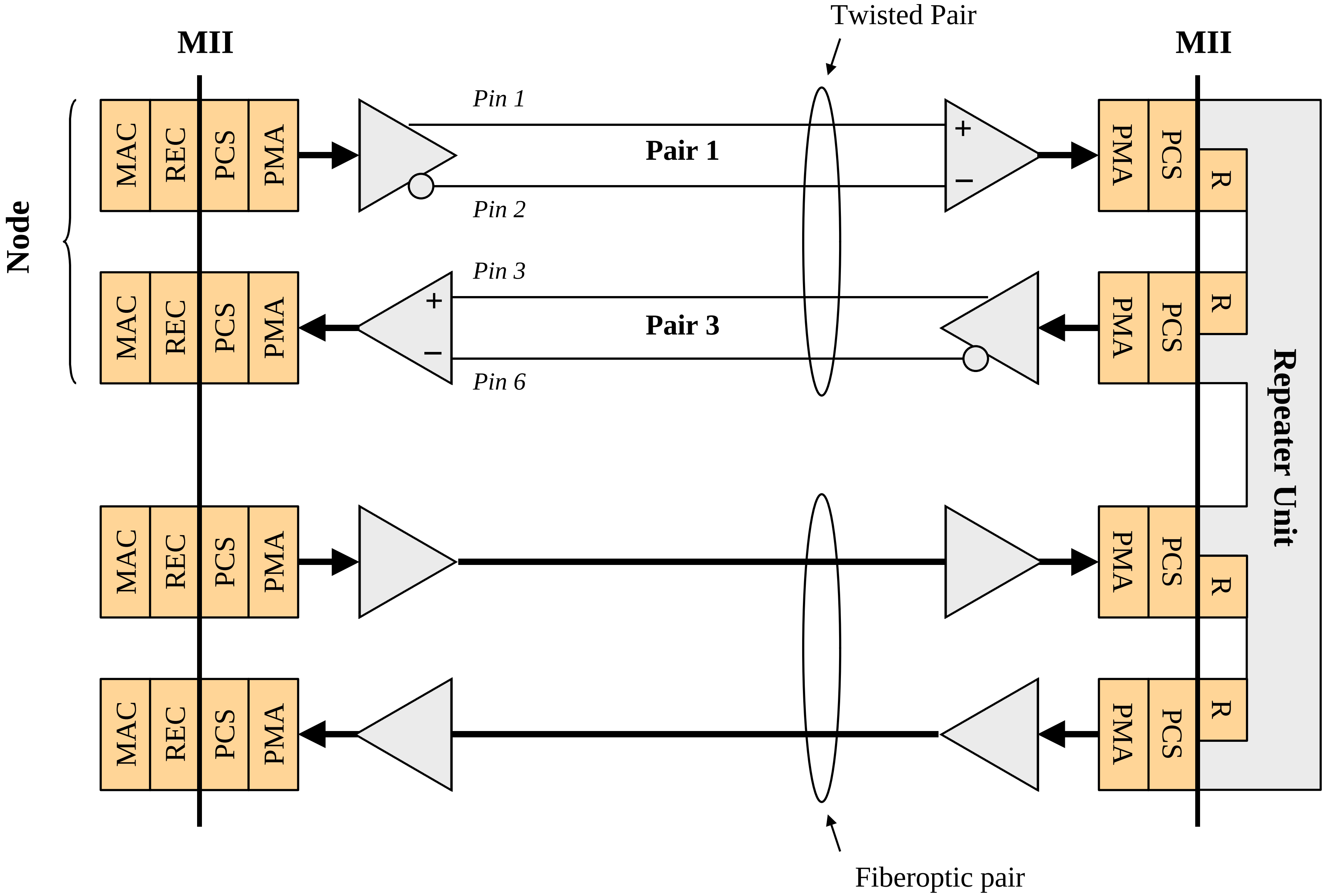
Missed Collision w/ oversized network



Node-to-node	**Path Delay Value** **(rounded to the nearest whole bit time)**
A → B	262 = 25 + 67 + 70 + 75 + 25
A → C	192 = 15 + 67 + 70 + 5 + 25
B → C	200 = 25 + 5 + 70 + 75 +25
A → B → A	524 = 2 * (25 + 67 + 70 + 75 + 25)
Safety Margin	4
Bit time margin	-16 = 512 -524 - 4

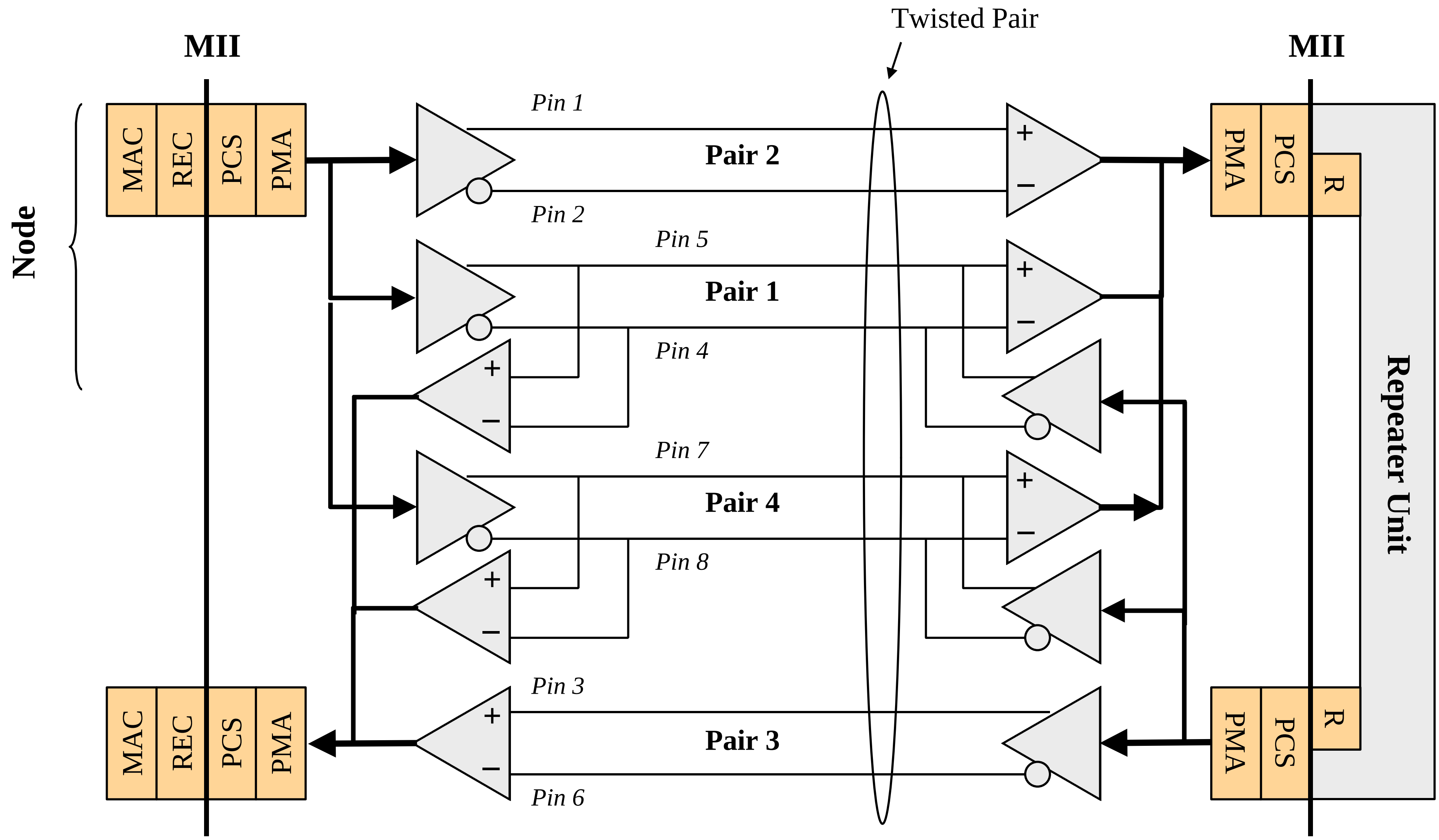


100Base-TX / FX Connection



ITRI
CCL

100Base-T4 Connection

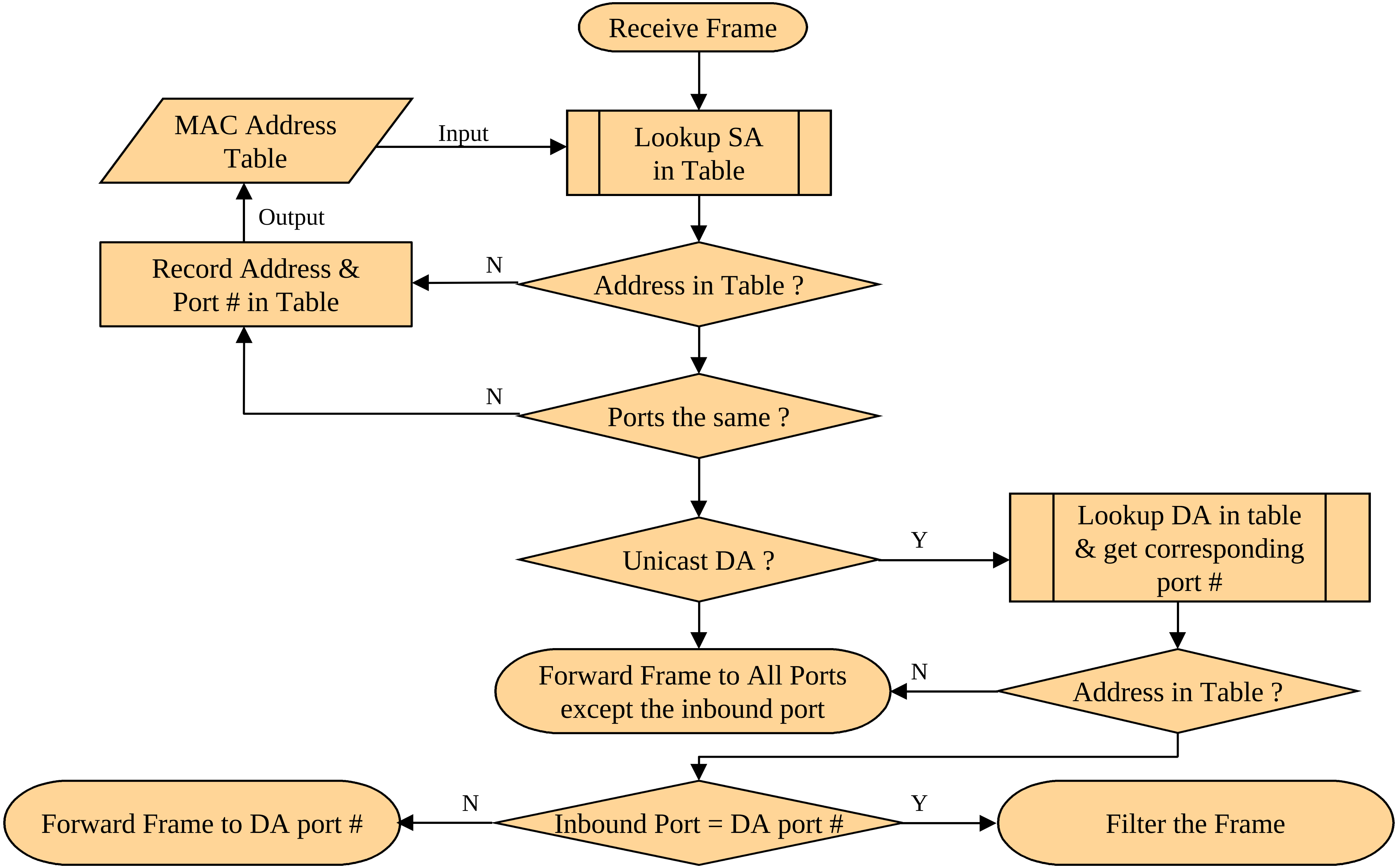


Repeater Testing

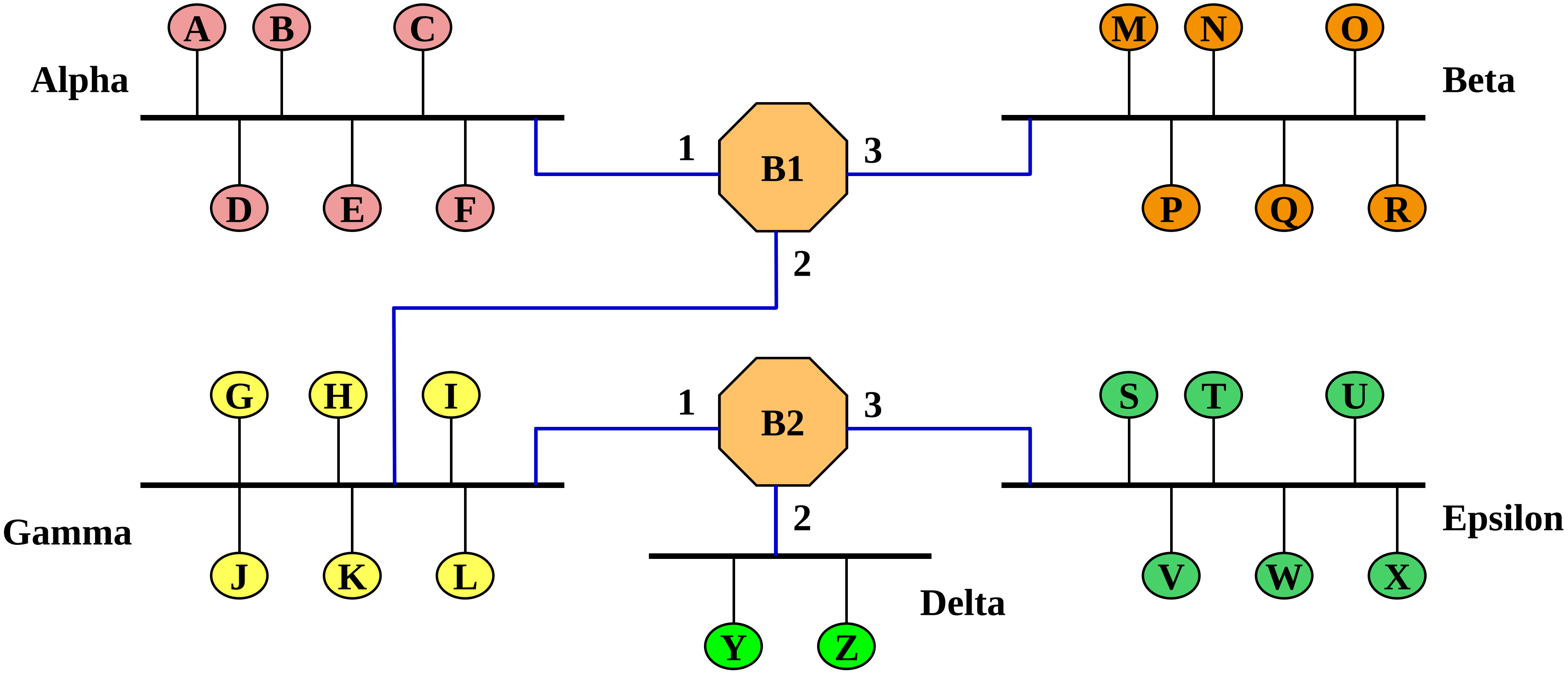
- Function
  - Transmit / Receive event
    - Data handling: forward packet
    - Receive event handling: carrier sense
  - Error handling via partition
    - False carrier events: invalid start-of-stream delimiter
      - Partition and set LINK UNSTABLE state after two false carrier event
      - Send jam signal to all other ports on the repeater for 5 µs or until end of FCE
      - Unset LINK UNSTABLE state after detecting no activity for more than 331 µs or detecting a valid incoming packet after the the line has been idle for the interpacket gap time of 640 µs.
    - Excessive collision: more than 60 collisions in a row
      - Partition after receiving more than 60 collisions in a row
      - Clear after detecting activity without a collision for more than 5 µs
    - Receiver Jabber: data transmission greater than 400 µs (largest packet: 120.56 µs)
      - Clear after jabber stops

ITRI
CCL

Basic Bridge Operation

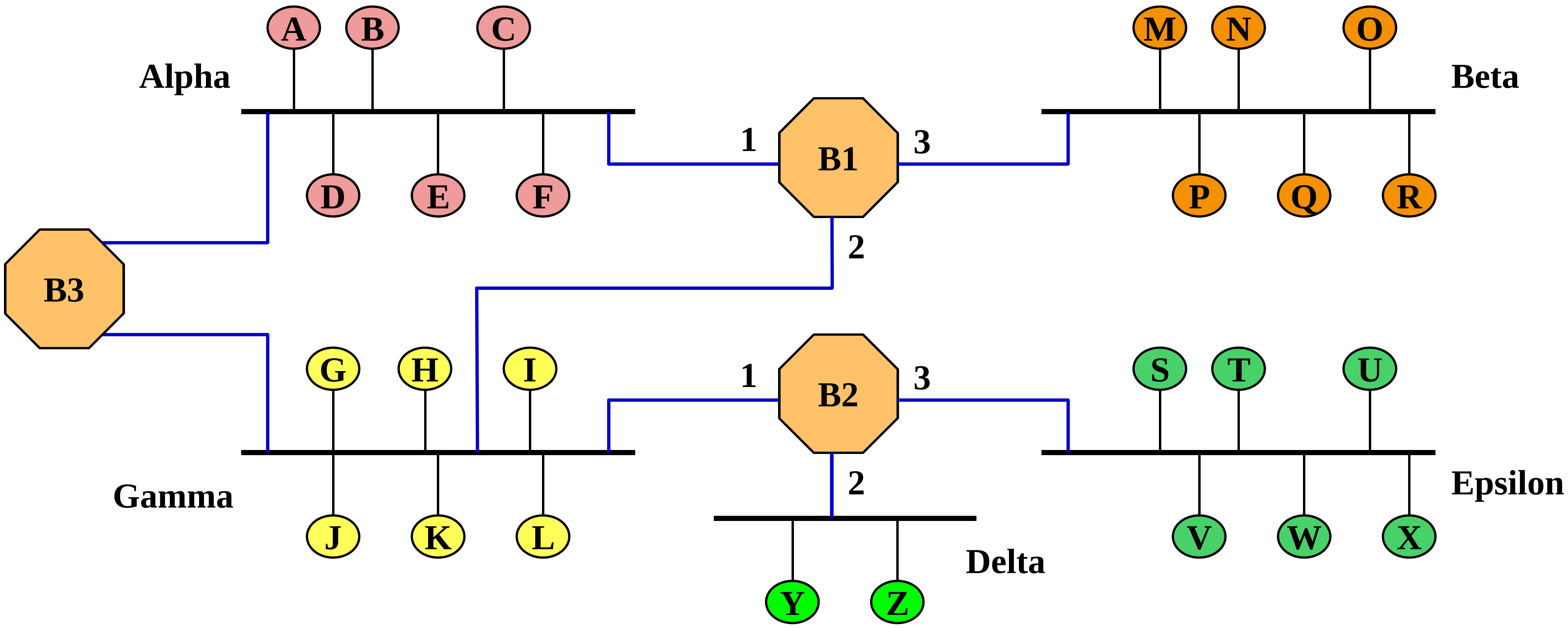


Multiple Bridges



Segment	Alpha	Gamma	Beta	Epsilon	Delta
MAC Address	ABCDEF	GHIJKL	MNOPQR	STUVWX	YZ
Bridge #1	111111	222222	333333	222222	22
Bridge #2	111111	111111	111111	333333	22

Multiple Bridges



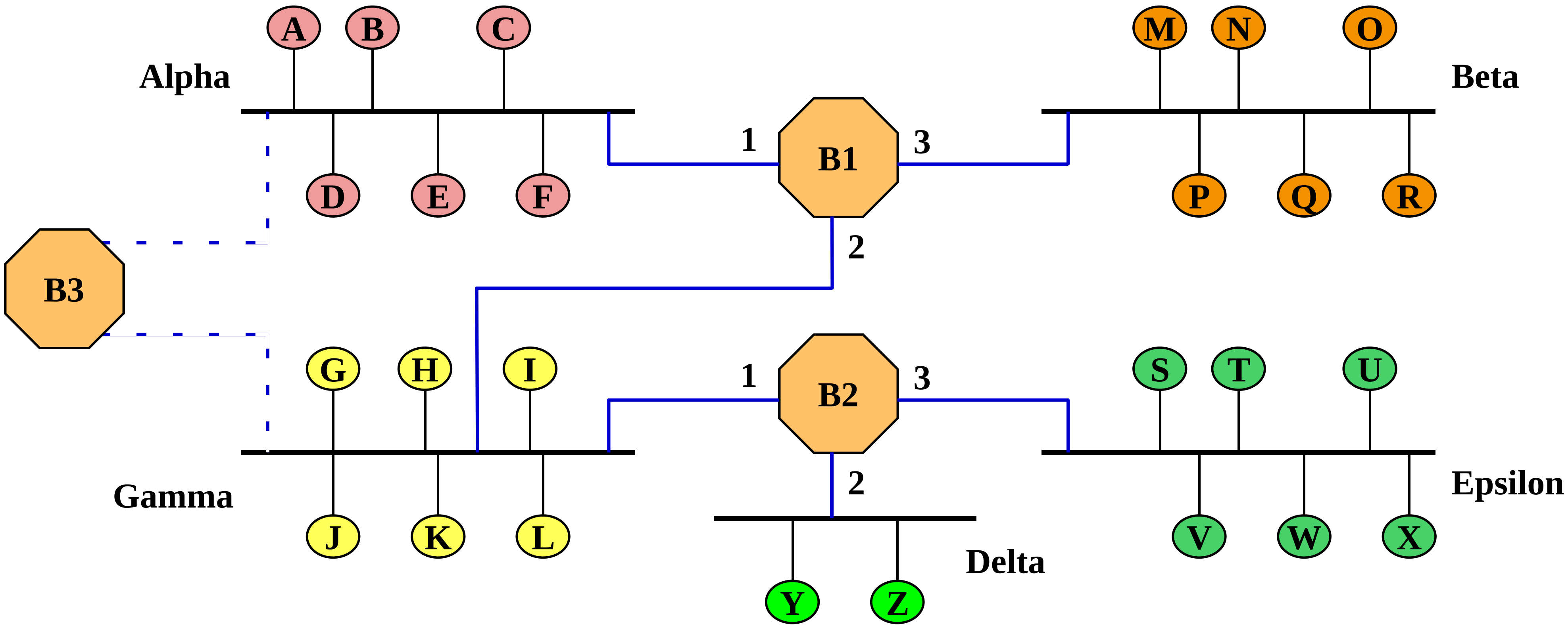
Problems caused by looping

- Broadcast storming
- Learning problems
- Cloned unicast frames

Solution

- Spanning Tree Protocol

Multiple Bridges



Problems caused by looping → **Solution**

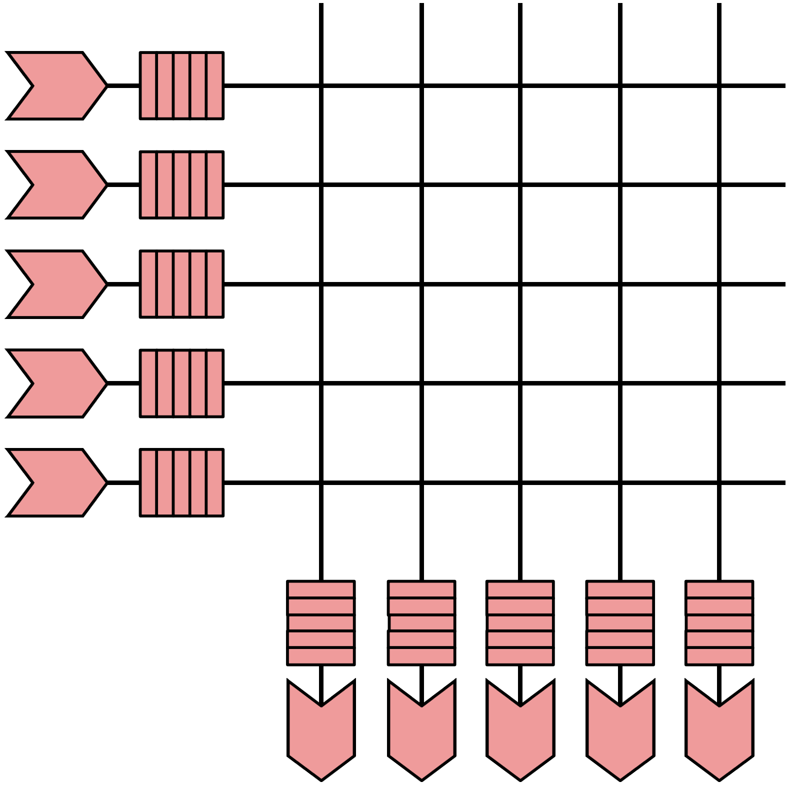
- Broadcast storming
- Learning problems
- Cloned unicast frames

Solution

- Spanning Tree Protocol

Ethernet Switching

- Basic techniques
  - Cut-through
    - Advantages
      - low latency
    - Disadvantages
      - forwards runt & error frames
      - internal speedup not possible
      - mixed speeds difficult
  - Interim Cut-through
    - Same as CT, but less runt frames
  - Store & Forward
    - Advantage
      - reduces error frames
      - architecturally flexible
    - Disadvantage
      - longer latency (not really bad !!)



Router vs. Routing function

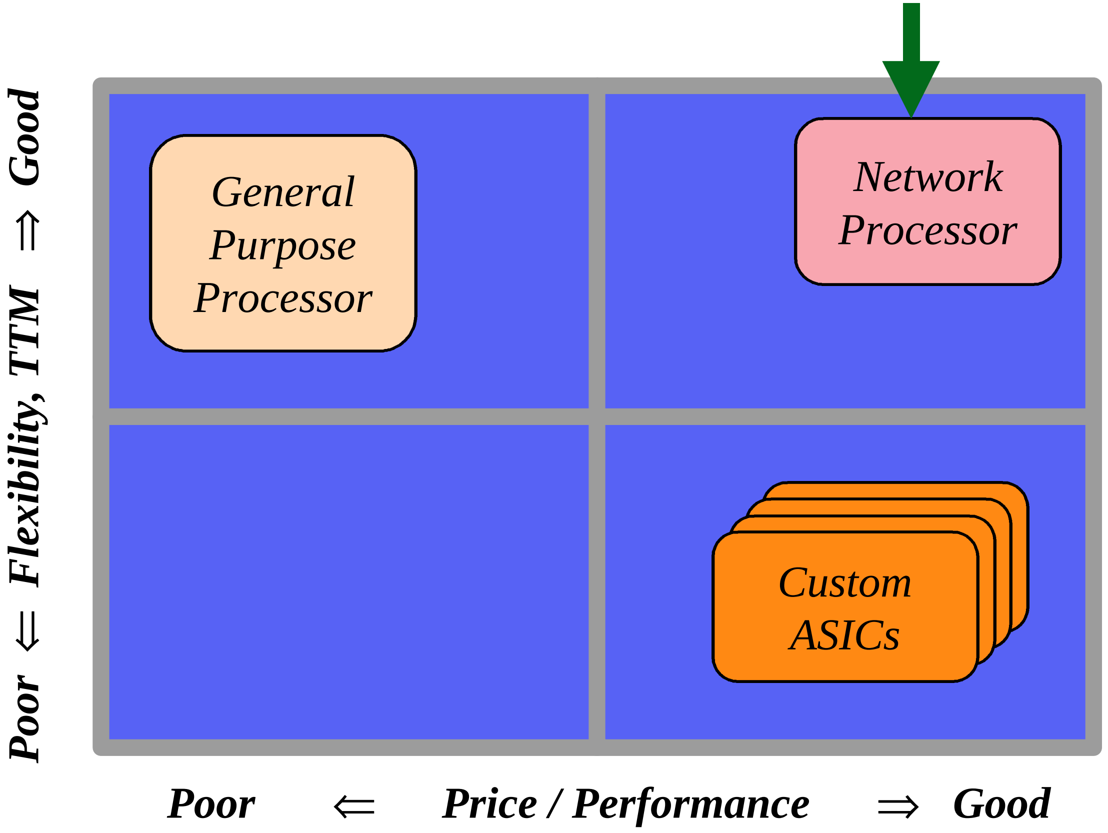
- What does a router do
  - » Routing function
    - – IP packet forwarding
    - – Route calculation/ convergence
    - – Route management
  - » Multicast
    - – IP packet duplication
    - – Multicast routing
  - » Traffic Mgt. (QoS)
    - – Packet Classification
    - – Packet Filtering
    - – Queue Management
  - » Network Mgt.
  - » Security
    - – Firewall
    - – Authentication

- Router
  - » Conventional stand-alone router performs an IP routing function
    - – Bus based
    - – Central CPU
    - – Cached forwarding tables
    - – Centralized routing tables
    - – SW table lookup
  - » Calculations required
    - – 10 Gbps throughput
    - – 64 byte packets = 50 ns / packet
    - – < 50 ns to make each routing decision.

Network Processor

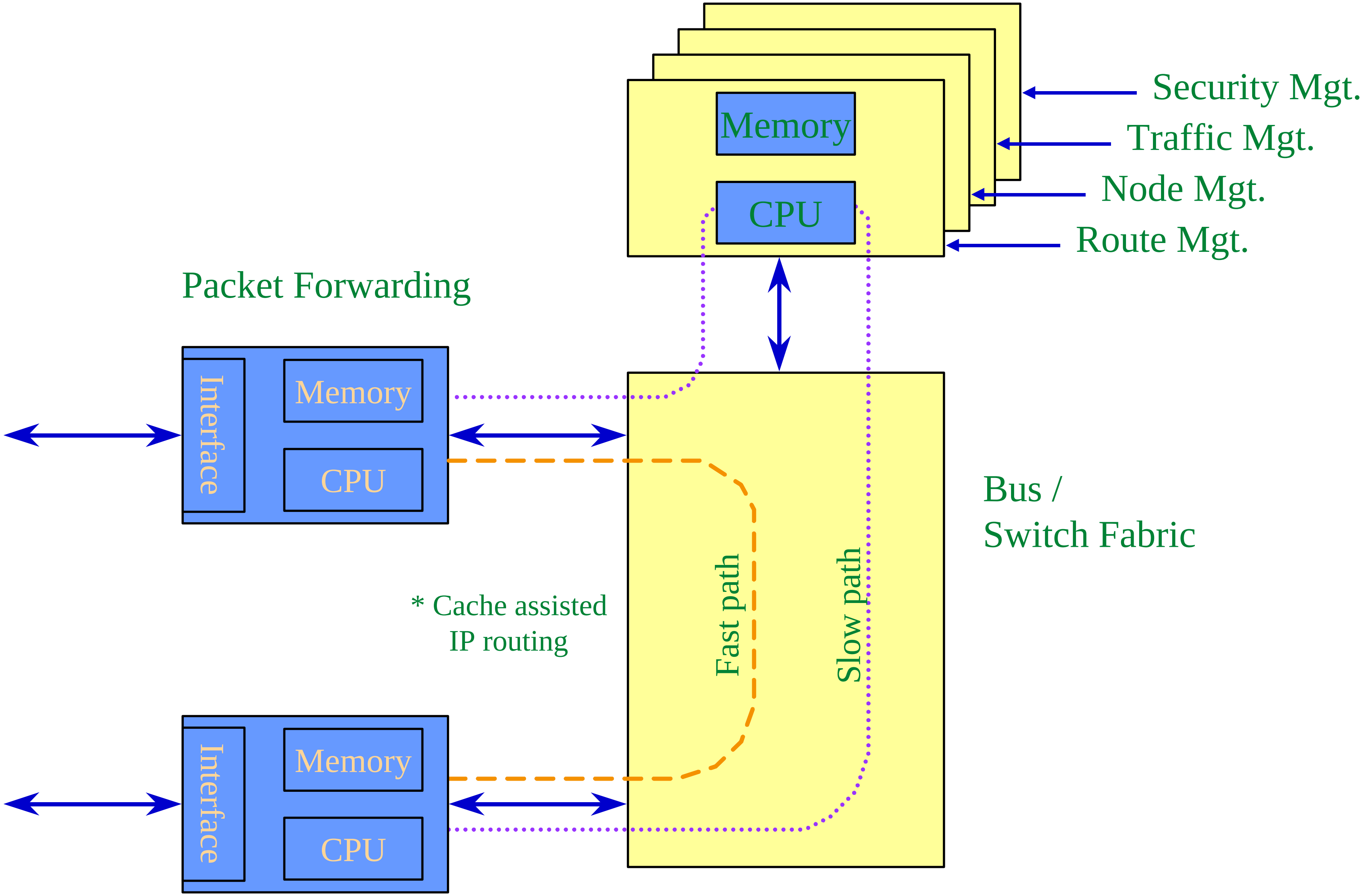
- Evolution to network processors
  - » Software programmable
  - » Optimized instruction set for networking
  - » Breakthrough performance
  - » Switching, routing, and features
- Customer-specific differentiation
  - » Base level instruction set
  - » Empowers the higher level software
  - » Addresses all networking markets
- Enables high-level functions at the same speed as basic switching wire speed

High Performance
Low Cost
Highly Flexible
Fast time-to-market


General Purpose Processor
Network Processor
Custom ASICs
Poor ⇐ Flexibility, TTM ⇒ Good
Poor ⇐ Price / Performance ⇒ Good

So, What’s so hard about switching & routing ?!#

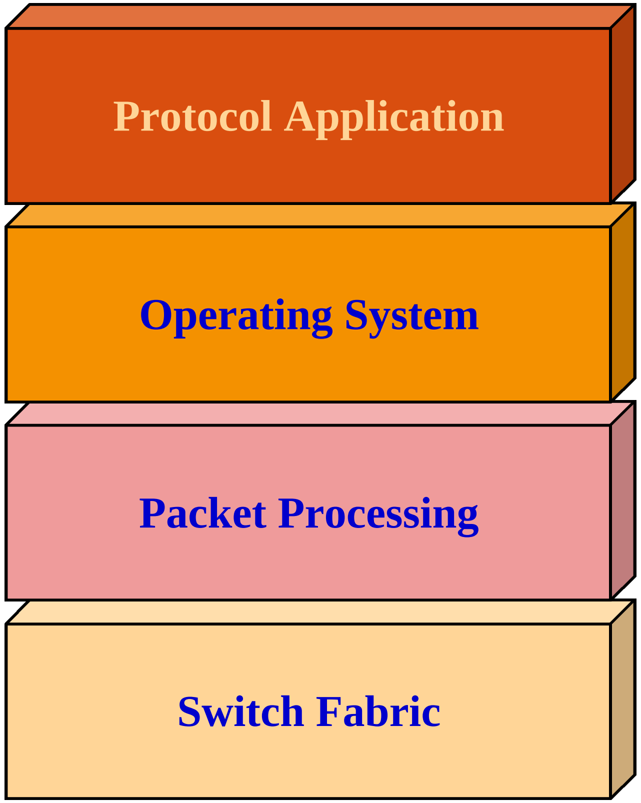
Typical Router Architecture



The Easy Part: Basic IP Forwarding

- To forward an IP Unicast packet, you need to:
  - » Parse the IP header
  - » Lookup the Address in a large table of address prefixes (100,000+ entries)
  - » Check the checksum
  - » Decrement the TTL and adjust the checksum
- This stuff is easy to do at high speed
  - » This is straightforward for ASIC implementation
  - » Clever implementers can do OC-192 (or even OC-768)
  - » With today’s technology, this is not even close to being the bottleneck

So What's So Hard?



Protocol Application

Operating System

Packet Processing

Switch Fabric

System performance = function of ALL elements
"A chain is only as strong as its weakest link"

So What's So Hard?

- There are things which can make high speed forwarding hard:
  - » Where data flows come together (backplane)
  - » Where parallelism is difficult
    - – e.g. Optics, software, protocol
  - » Protocol standards
    - – Unstable or poorly designed or under-defined standards
    - – Need mature implementations
    - – Multi-lingual
    - – Too many standards
  - » Lots of options and alternative paths
  - » Maintaining per-packet state that comes and goes

So What’s So Hard?

- Proliferation of standards make system implementation hard:
  - Support for legacy protocol (i.e. Multi-protocol & conformance)
  - Interoperability (i.e. Multi-vendor)
  - Addressing
  - Routing
  - Multicasting
  - Traffic mgt. (QoS)
  - Network mgt.
  - Mobility
  - Security
  - Virtual Private Network

So What’s So Hard?

- Reliability, maintainability, redundancy
  - » Hot swappable, Hot standby router
  - » Coherent network state
  - » Online upgrade
  - » Redundancy (power supply, link failure, etc.)
- Scalability
- Additional
  - » Frame translation
  - » Load balancing
  - » Port mirroring

One Hard Part: The Backplane

- Given 100 OC-xx ports, served by 100 line cards, somehow packets have to get between the line cards
- The design of the switched backplane is “non- trivial
  - » If there are “n” line cards, you have an n*log(n) problem
  - » You want very high switch utilization to push performance (this effects where packet is buffered)
  - » Power and heat become important
  - » Cost of hardware is meaningful
- It is easy to lose your QoS guarantees across the switched backplane

ASIC Designer's Nightmare: Options

- MPLS and IP forwarding
- Filters (source or destination address, RSVP, …)
- Tunneling: Encapsulation and decapsulation (particularly if reassembly is needed)
- Multicast
- IP Options
- Multipath (ECMP)
- NAT (application addresses plus state)
- IPv6 alternate headers

What does MPLS do to a router ?

- Answer: Provide lots of alternative forwarding paths

<IP>	→	IP
<IP>	→	<Shim> + <IP>
<IP>	→	<ATM+ shim> + <IP>
<Shim>	→	<Shim> ; <Shim>+<Shim>; <ATM+ shim>
<Shim>+< IP>	→	<IP> (with or without IP lookup)
<ATM+ shim>	→	<ATM+ shim>; <Shim>;…
<ATM+ shim>+< IP>	→	<IP> (with our without IP lookup)
<ATM+ shim>+< Shim>	→	<Shim>

etc …

- This is not popular with hardware developers ☹

A Hardware Developer’s View

- Generally: Hardware engineers wish that folks who write standards paid attention to hardware issues
- IP Forwarding can be done very fast, no problem
- CLNP forwarding can be done very fast, …
- But: Please don’t give us so many options!
- “It is clear that IP standards (including IPv6) were designed by folks who don’t pay any attention to what it takes to build a fast router?”
- (On the other hand, things are still within a top hardware team’s capabilities)

The Bottom Line on Speed

- There are really three bottlenecks:
  - » The switched backplane
  - » The optics
  - » How much extra complexity and flexibility you want (Filtering, MPLS, options, all make it harder to go fast
- It really doesn't matter what the forwarding looks like, if its straightforward and well defined
- At very high speed, IP, MPLS, ATM, Frame Relay, are all constrained by the same issues are all constrained by the same issues

Can Routers go fast enough ?

- There is some limit to how fast routers can go
- Or, more correctly, there is some limit on how fast electronics can go
  - Given today’s chip technology, and reasonable economics, the limit might be on the order of a few thousand * OC-192
  - In four years, possibly ditto but * OC-768
- Past this point, we need optics
  - Core switches become WDM switches
  - Very fast, very branchy routers (and ATM switches) become feeders for WDM in the core

Issues affecting Reliability

- Hardware robustness
  - » Reliable hardware, Redundancy at many levels
- Software quality and robustness
  - » e. g., How good is your routing software?
- Protocol Design Protocol Design
- Response to congestion Response to congestion
- Failover of links (“Sonet- Like” failover rates)
- Network Management
- Avoid mistakes, Diagnose failures
- Testing, testing, testing

Key Design Considerations

Are your assumptions reasonable ?

Design Constraints

- Target: *Right product at the right time at the right price*
  - » Cost
  - » System
  - » Market Segment
  - » Market Timing (Market window)
  - » Specifications
- Competition
  - » Advantages & Weaknesses
  - » Targeted Market
- Resources
  - » Engineering team (Experience / Stability)
  - » Management team (Financing / Supportiveness)
  - » Standards / Customer / Industry tracking

What is Required ??? (Perceived vs. Real)

- Optimizing Performance:
  - Wire-speed switching at Layer 2
  - Wire-speed forwarding at Layer 3
- Minimize Latency: Cut-through switching vs. Store-and-forward
- Increased Scalability: SOHO, Departmental, Enterprise, Backbone
- Maximize Integration: Multi-chip vs. Single chip solution
- Increased Functionality:
  - VLAN (Port, MAC, IP, IEEE 802.1q Tagging, etc.)
  - Port Trunking / Port Snooping
  - Support Layer 3, Layer 4, … , Layer 7
  - Support IP, IPX, SNA, …
  - Support IEEE 802.3x flow control, jamming
  - CoS / QoS / RSVP / SBM / Differentiated Service
    - Multiple loss / delay queues
    - per VC queueing

Are these assumptions reasonable ?

- Maintain multicast / unicast packet sequence.
- Multicast packet needs to switched at the same time
- Support 8 k / 16 k / 32 k MAC addresses.
- Support 8 k / 16 k / 32 k IP addresses.
- Support full SNMP / RMON statistic collection.

Key Design Considerations

Can you successfully overcome today's technology limitations ?

Technology Assumptions

- Memory speeds, size, types
  - » DRAM, SRAM, SDRAM, SSRAM, Rambus, NetRAM
- Semiconductor technology
  - » Dimension: 0.8 µm, … , 0.35 µm, 0.25 µm, 0.18 µm, etc.
  - » Power: 5 V, 3.3 V, 2.5 V, etc.
  - » Embedded Memory
- Design Tools
  - » Simulation: RTL level, Behavioral, Cycle-base
  - » Layout Capabilities
  - » Emulation Technology

Do these assumptions hold ?

- Freebies
  - » Memory speed / size
  - » Silicon cost
  - » Computational power
- Does these assumption still holds in a hyper-competitive environment?
  - » No, because everybody have access to the same components and semiconductor foundry.

Improved Competitiveness

- Creating advantages by speeding up design cycle
  - Increase engineering experience
  - Invest in the state-of-the-art engineering tools
    - Latest simulation / CAD tools
    - Advance computers
    - SOC Emulation / FPGA hardware
      - to improve simulation time
      - to reduce potential errors, thus less debugging

Design Goals

Design specifications

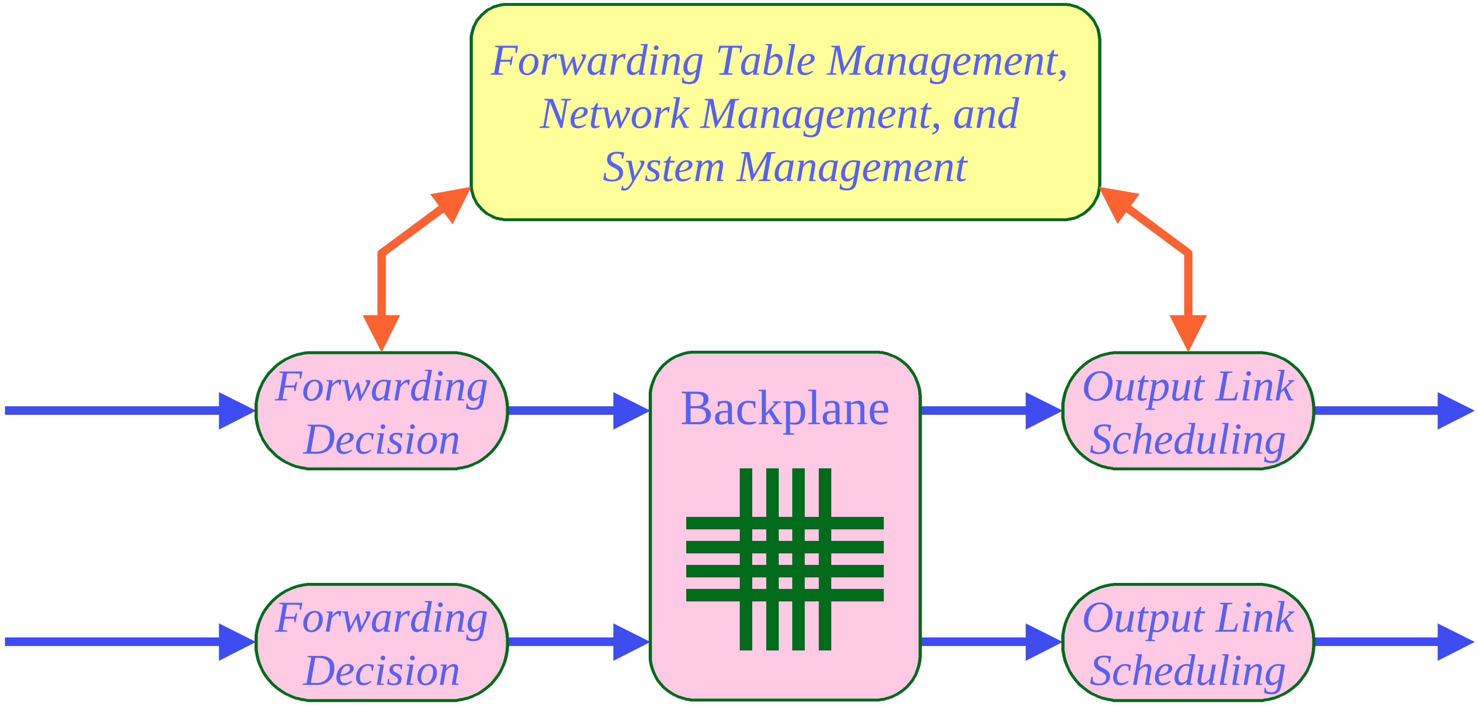
Design Specifications

- System Features
  - » Single chip eight 10/100 Mbps Ethernet ports with RMII interface
  - » Provides two 32-bit memory interfaces which support SSRAM
  - » Supports a 16-bit CPU interface
  - » Statistics collection to support SNMP, RMON-1
- Layer 3 Features
  - » Supports wire-speed IP routing (1.2Mpps) with line rate address lookup
  - » Supports 10K routes
  - » Supports IP Multicast
  - » Supports two level of user data priority (Class of Service Support)
- Layer 2 Features
  - » Supports IEEE 802.1d bridging and spanning tree algorithm
  - » Supports port or IEEE802.1Q compliant tag based VLANs
  - » Supports 8K MAC address entries
  - » IEEE 802.3x flow control for full duplex operation
  - » Supports port snooping

Design Goals

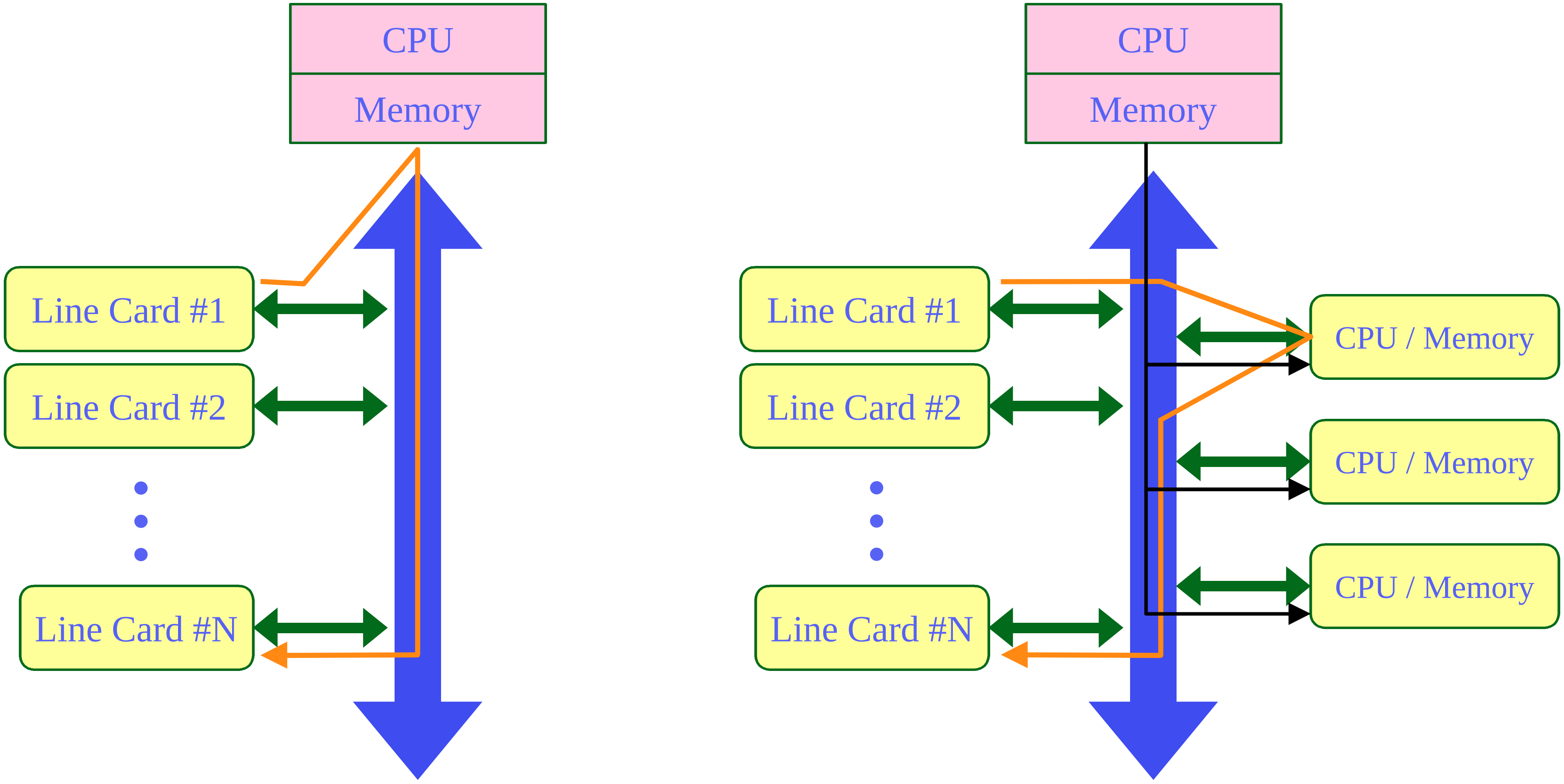
Architecture

Key Common Architectural Components

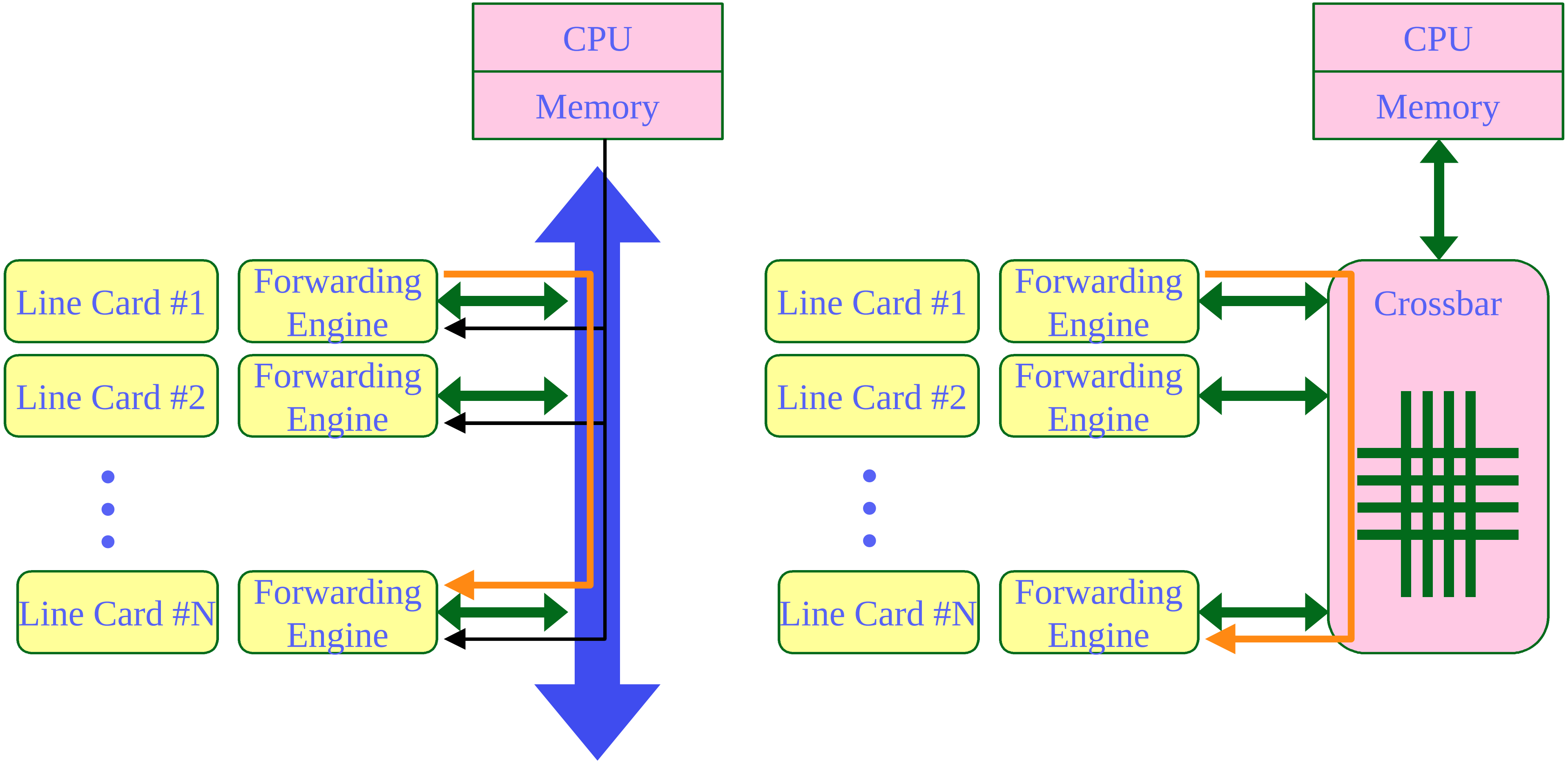


In high performance systems, the forwarding decision, backplane and output link scheduling must be performed in hardware, while the less timely management and maintenance functions are performed in software.

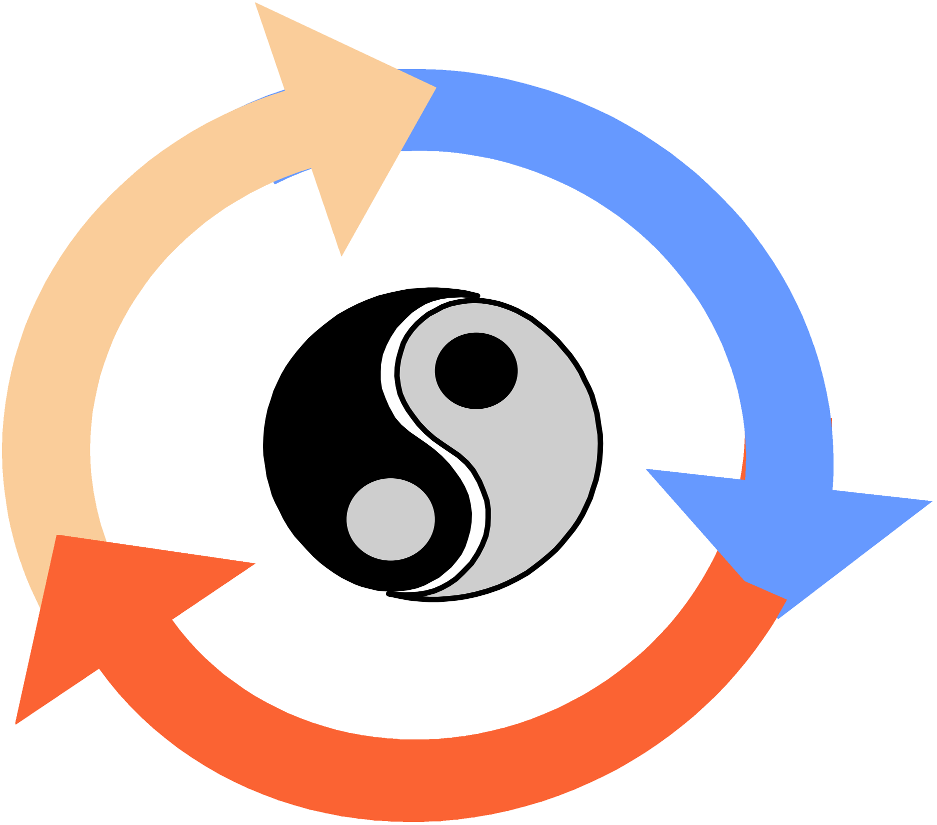
Architectural Evolution



Architectural Evolution



Architectural Renewal w/ Advance Technology



Trade-off

Centralized	vs.	Distributed
Hardware	vs.	Software
Packet	vs.	Cell
Connection-less	vs.	Connection-oriented
Shared Bus	vs.	Crossbar
Transmission	vs.	Switching
“Big Pipe”	vs.	“Managed BW”
“Dumb Network”	vs.	“Intelligent”
Wired	vs.	Wireless

Technology Factors

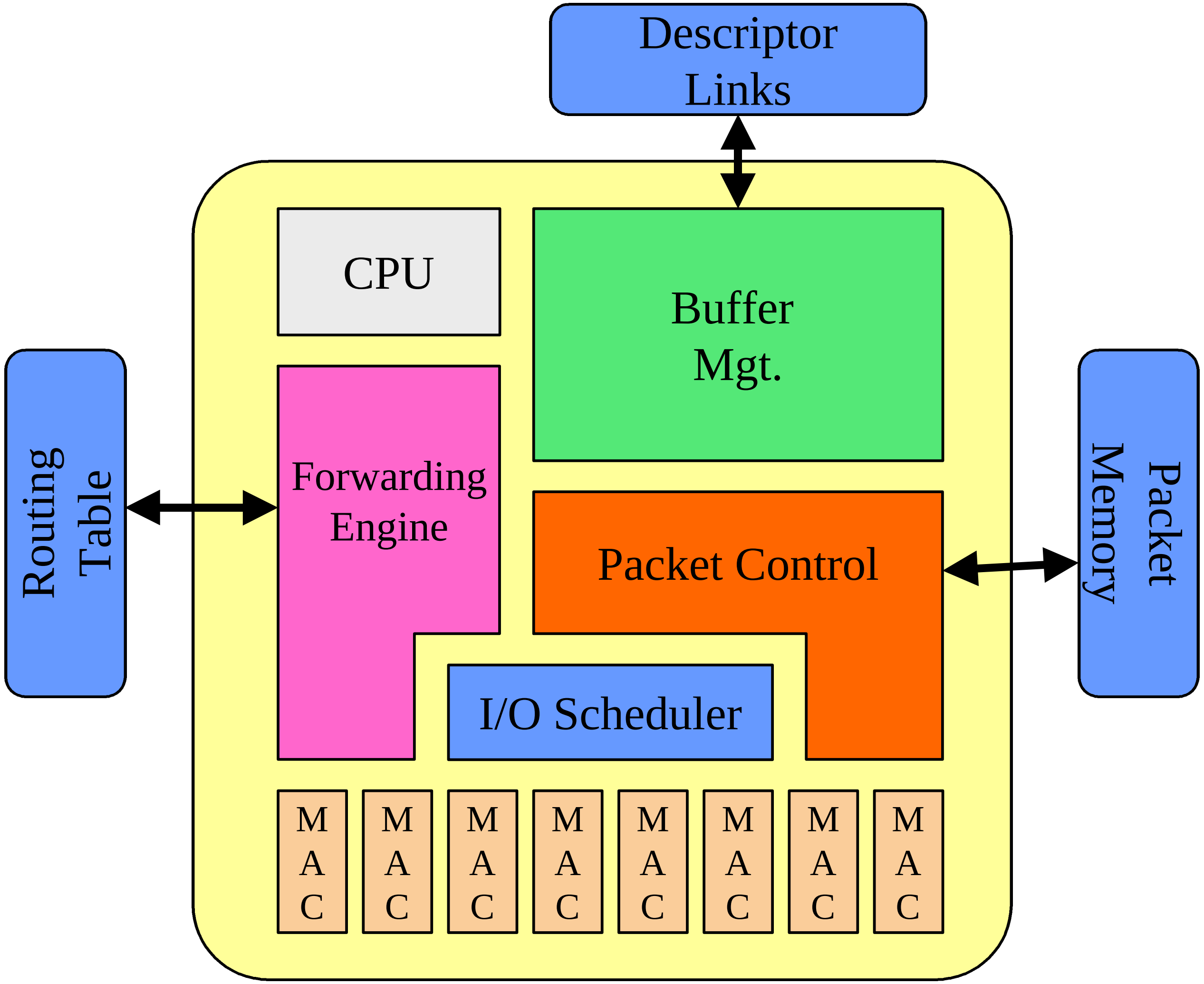
Semiconductor Advances

- Computing Power (CPU)
- Memory Size
- Analog / RF / Optical technology

Material Advances

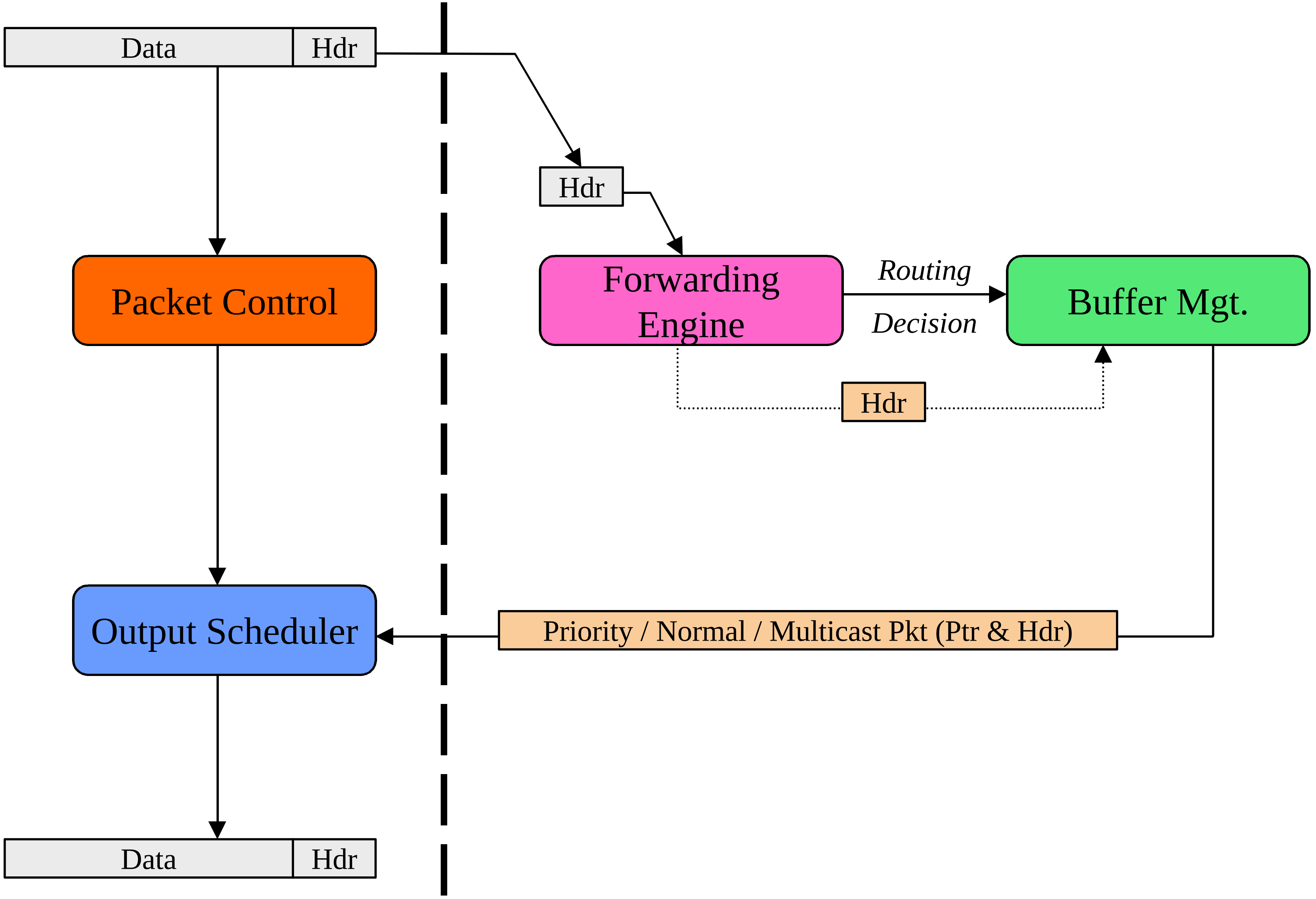
- Optical transmission

Conceptual Model of L3 Switch

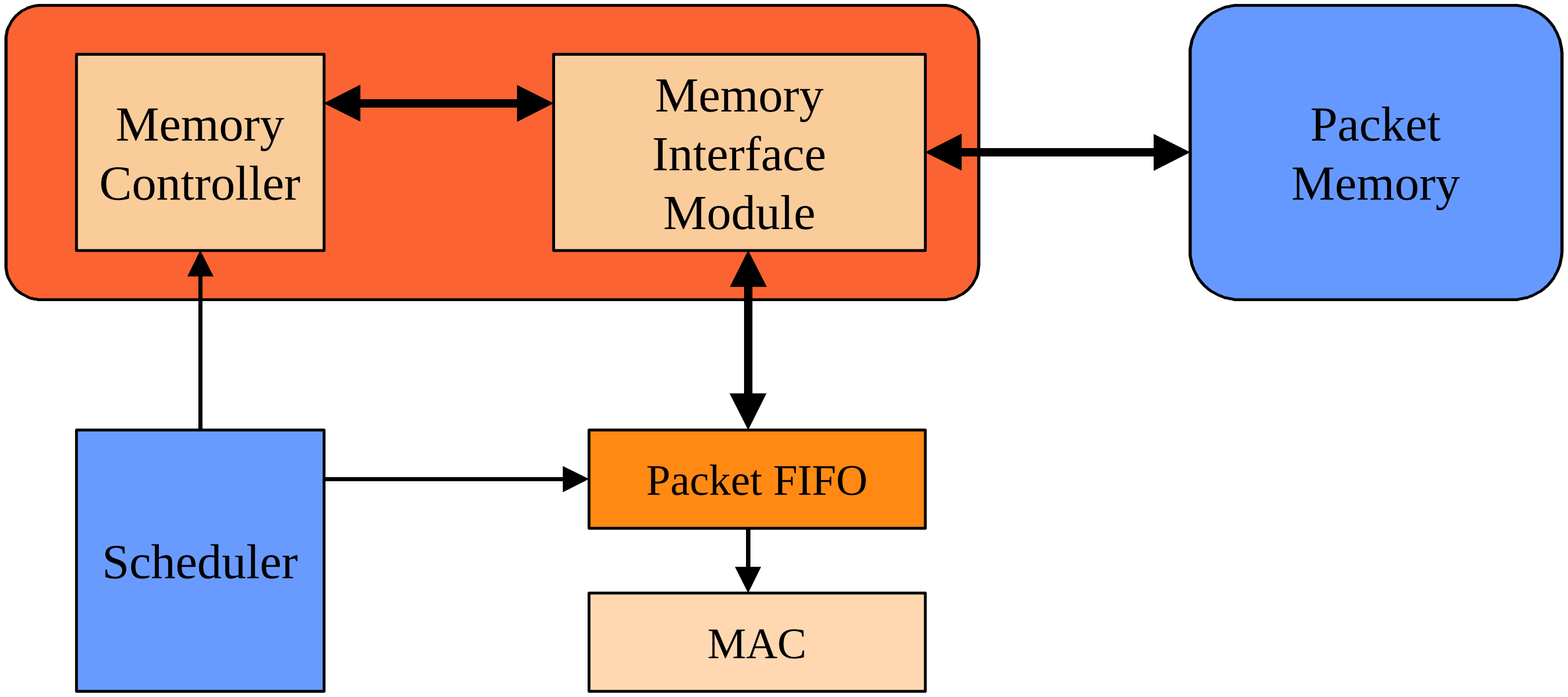


ITRI
CCL

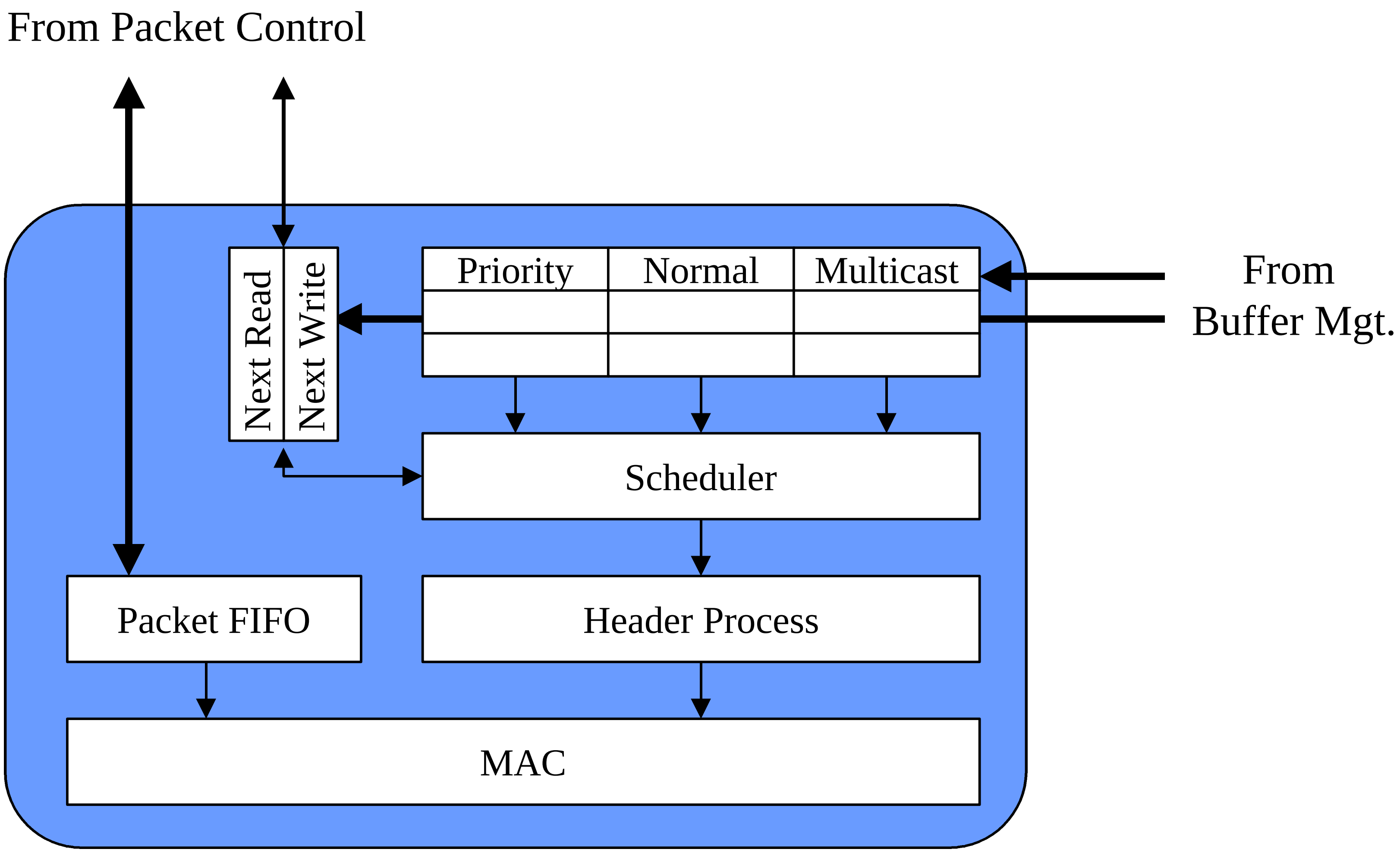
Packet Flow (Tetris)



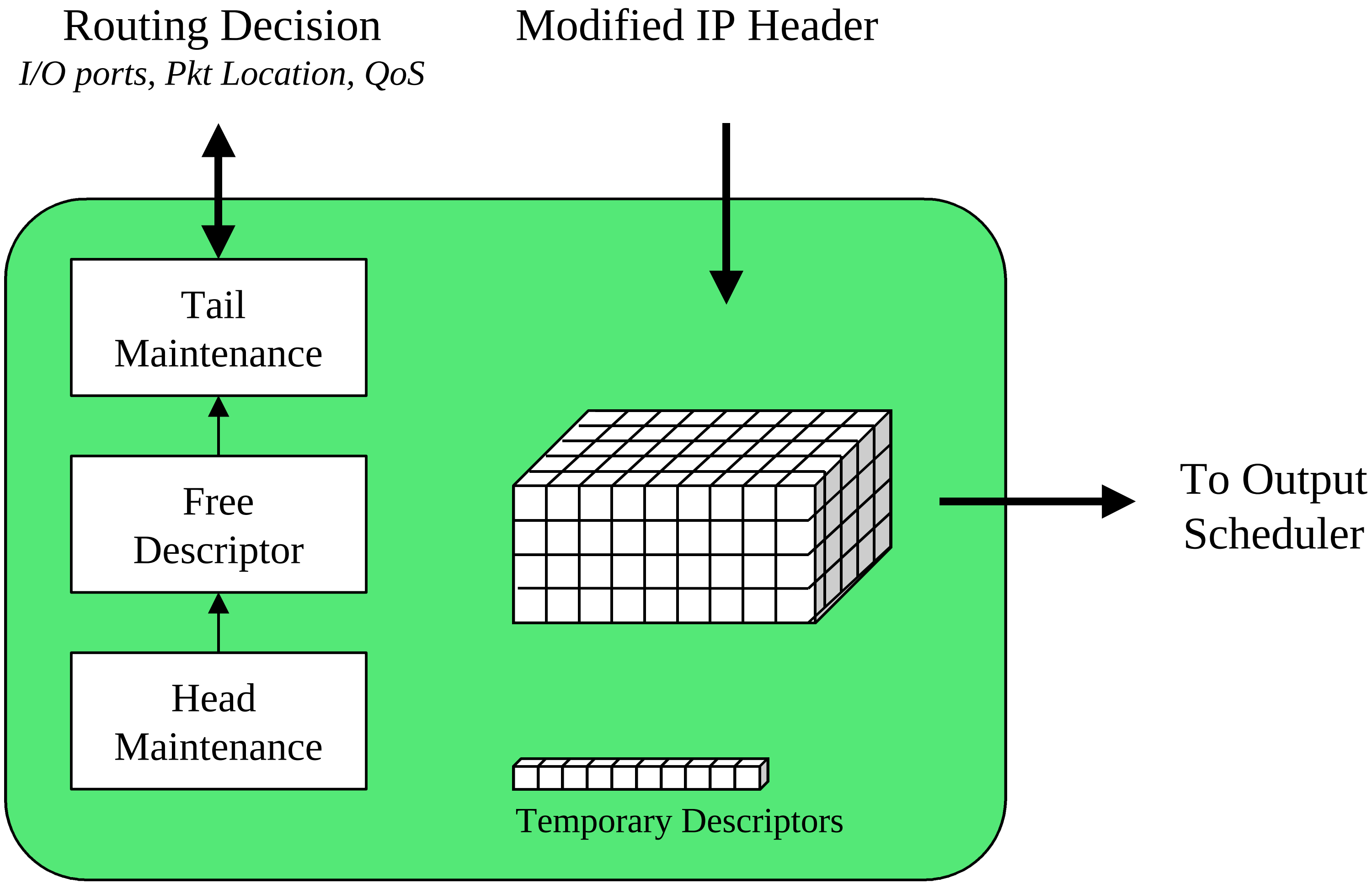
Packet Controller



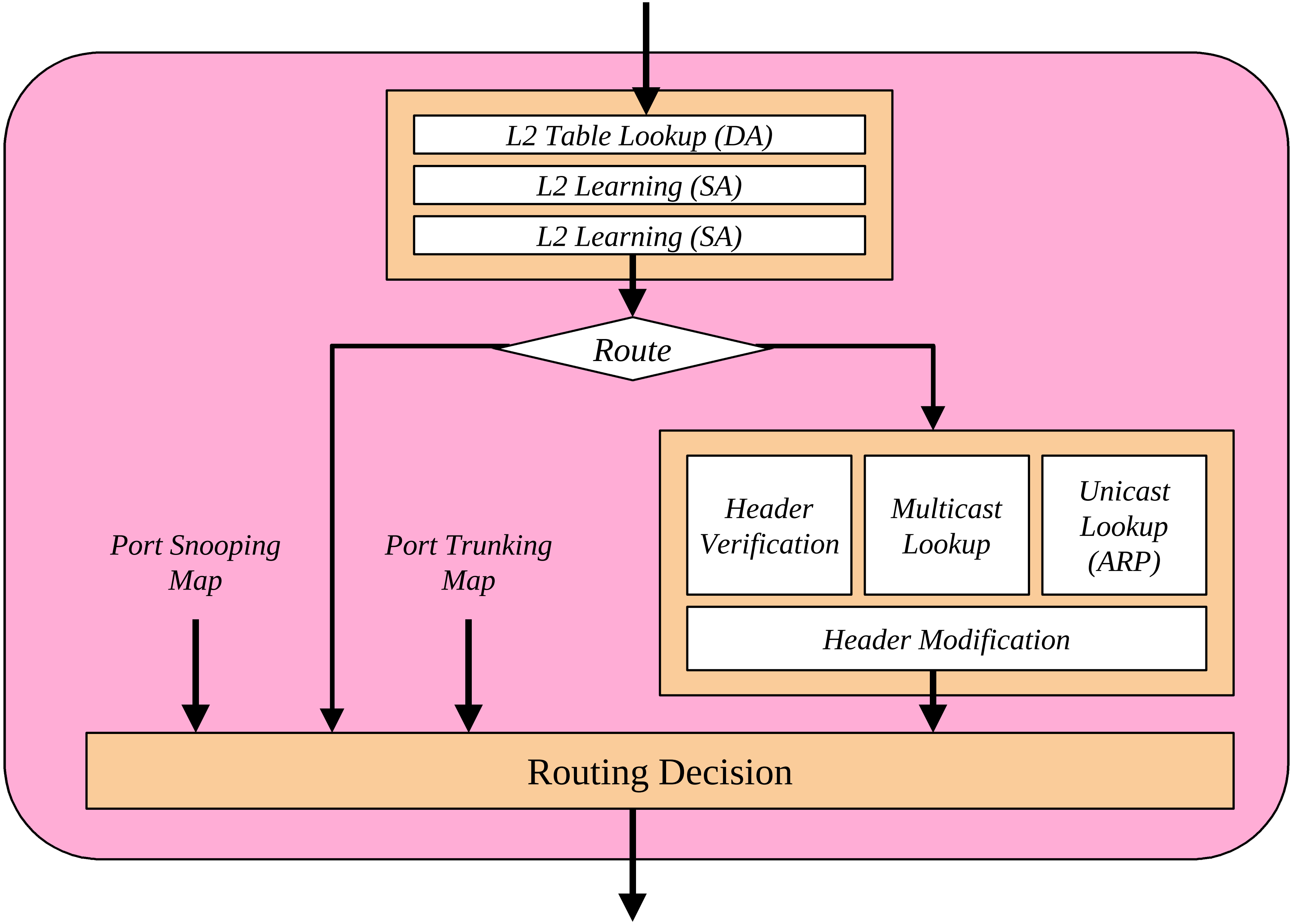
Output Scheduler



Buffer Management



Forwarding Engine



L2 Table Lookup (DA)

L2 Learning (SA)

L2 Learning (SA)

Route

Header Verification

Multicast Lookup

Unicast Lookup (ARP)

Header Modification

Port Snooping Map

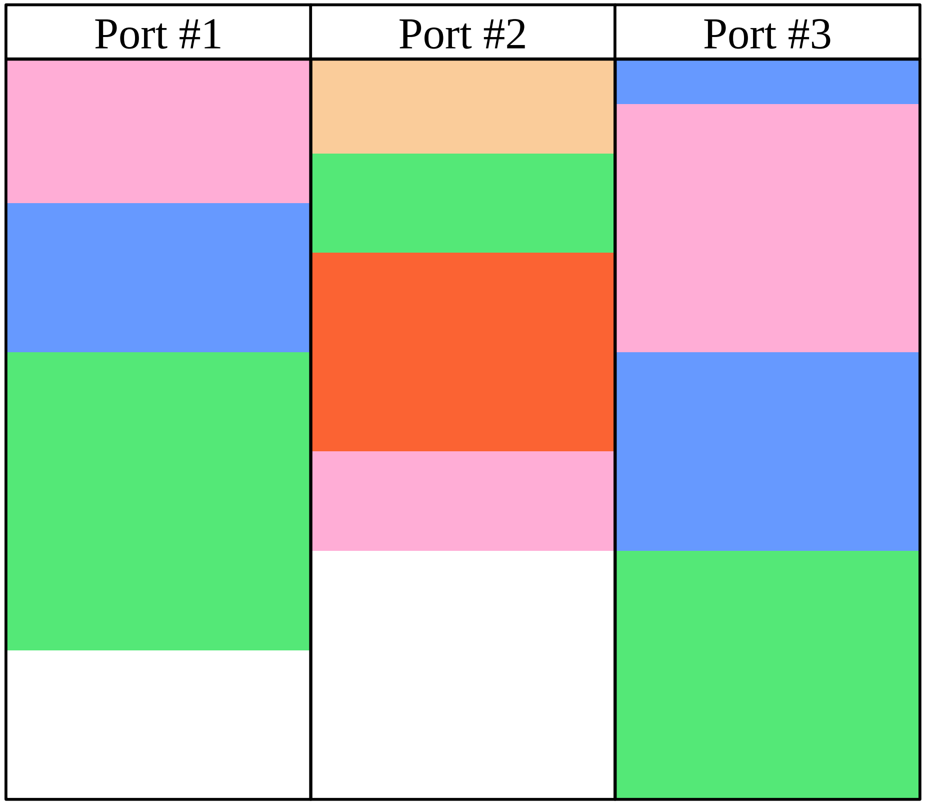
Port Trunking Map

Routing Decision

Design Trade-offs

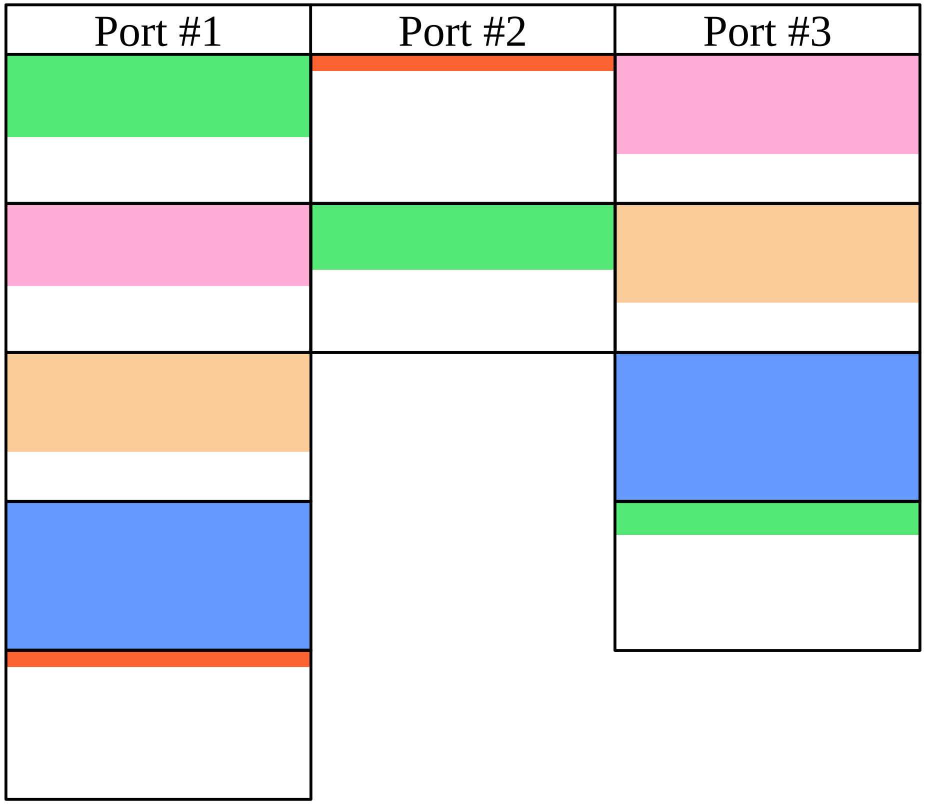
Packet Memory Design

Variable Length Format



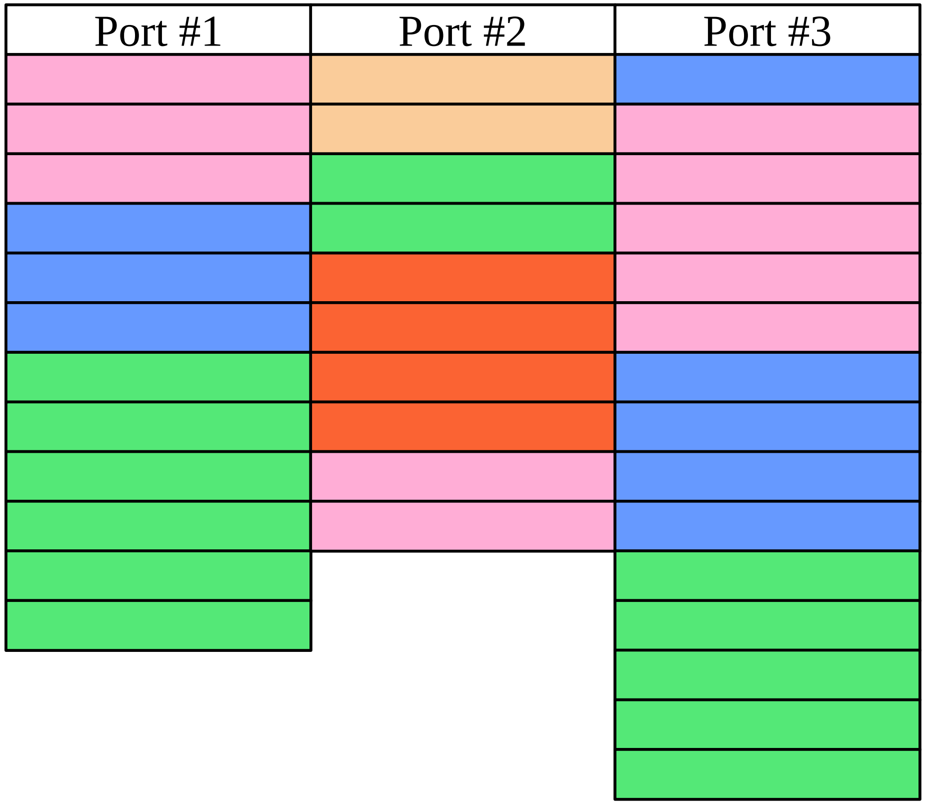
- Advantage
  - » Simple, no link list required
  - » No descriptors required
  - » No gaps between packets
  - » Easy to debug
- Disadvantage
  - » No sharing among ports
  - » Fast route decision required
  - » Large temporary FIFO required
  - » Parity bit or packet length write-back required
  - » Look-ahead forwarding not allowed for multicast packets
- Variations
  - » Parity bit vs. Packet Length

Fixed Packet Size Format



- Advantage
  - Sharing among ports
  - Routing decision relaxed
  - Look-ahead forwarding allowed
  - Small temporary FIFO
- Disadvantage
  - Inefficient for small packets
  - Link list required
  - Difficult to debug
- Variations
  - 1536 bytes vs 2048 bytes

Cell Format

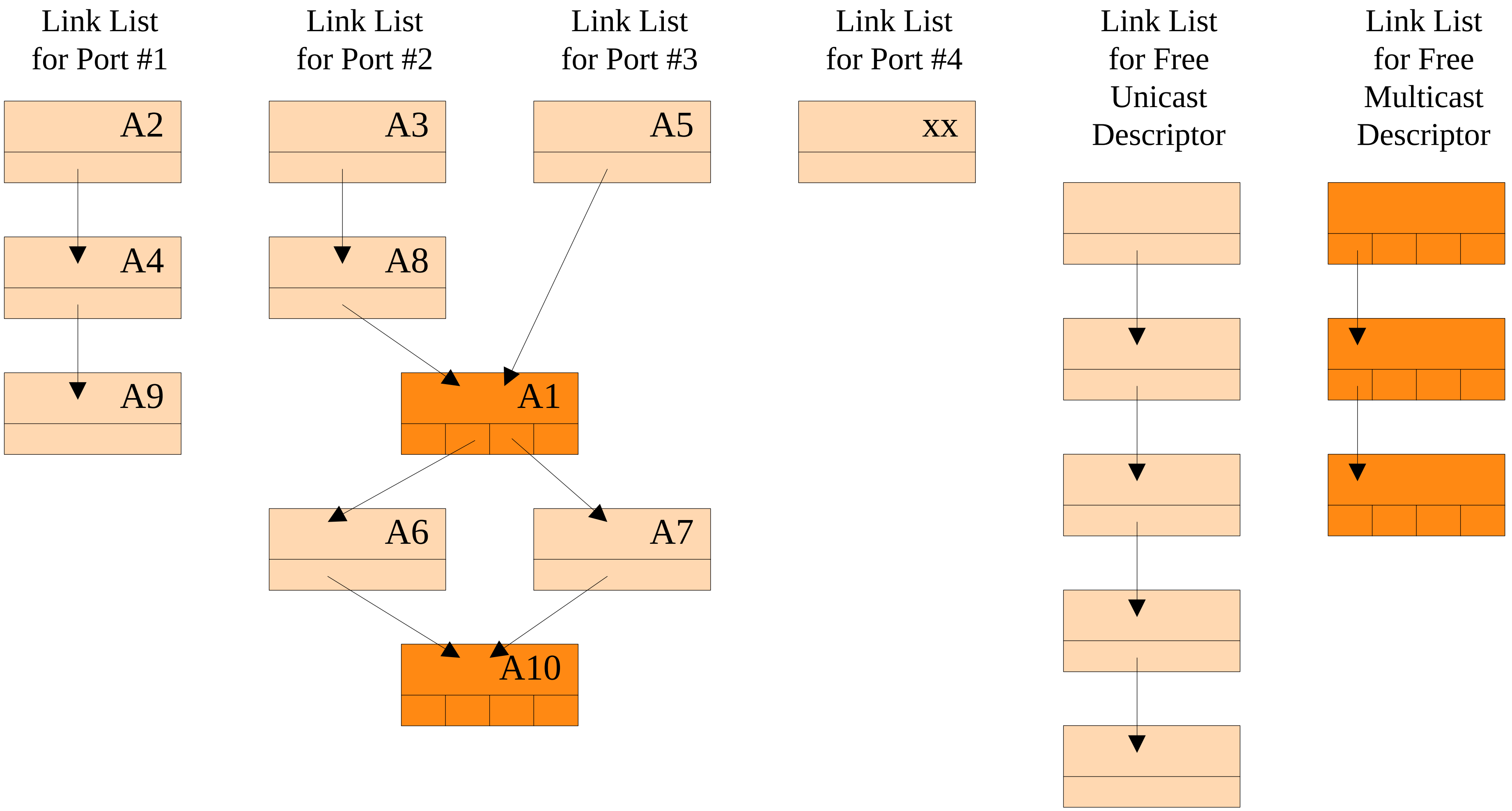


- Advantage
  - » Sharing among ports
  - » Efficient for most packets
  - » Routing decision relaxed
  - » Look-ahead forwarding allowed
  - » Small temporary FIFO
- Disadvantage
  - » Large descriptor memory required
  - » Link list required
  - » Complex logic / Longer design cycle
  - » Prone to error
  - » Very difficult to debug
- Variations
  - » 64 / 128 / 256 bytes

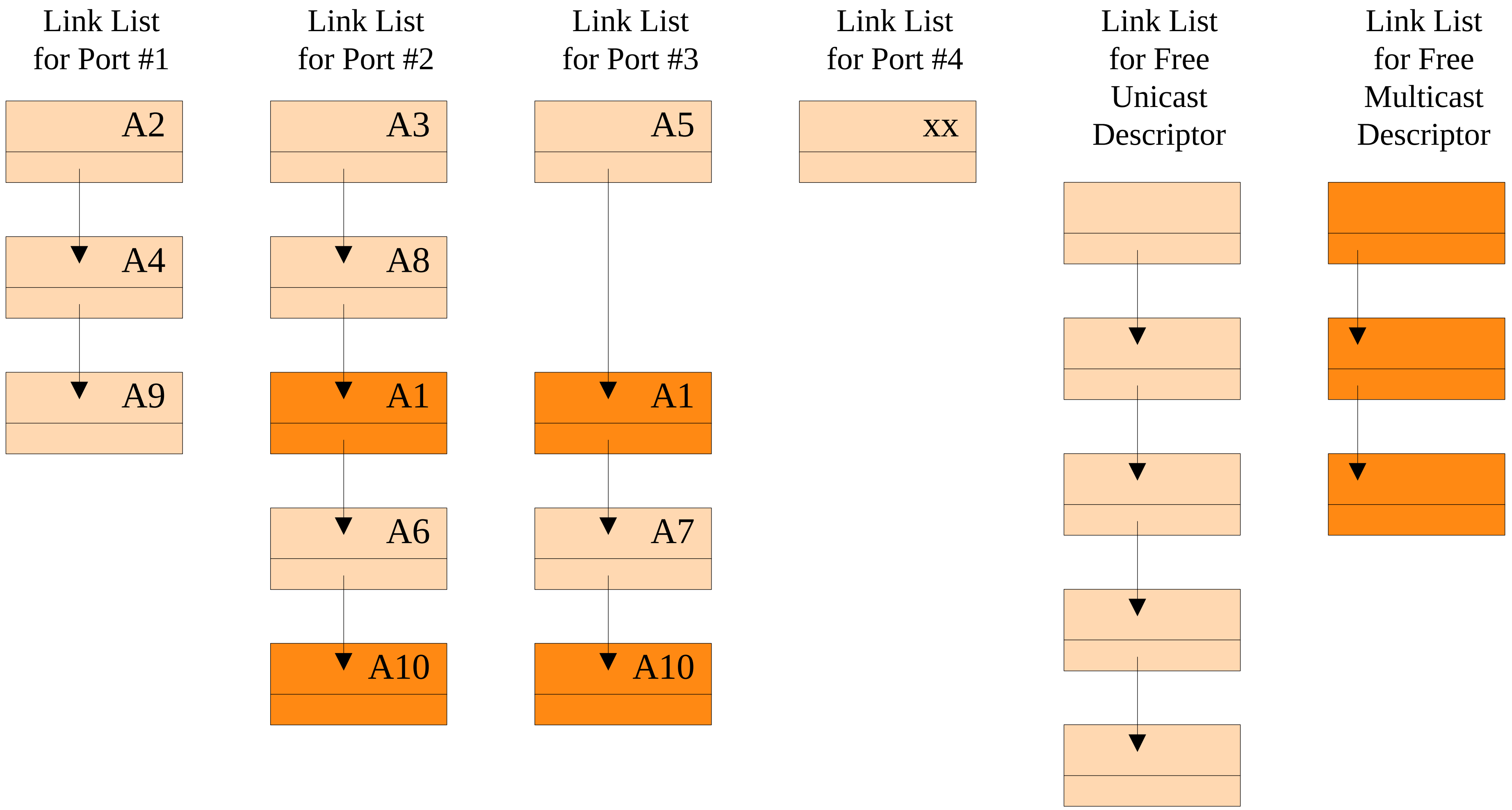
Design Trade-offs

Buffer Management

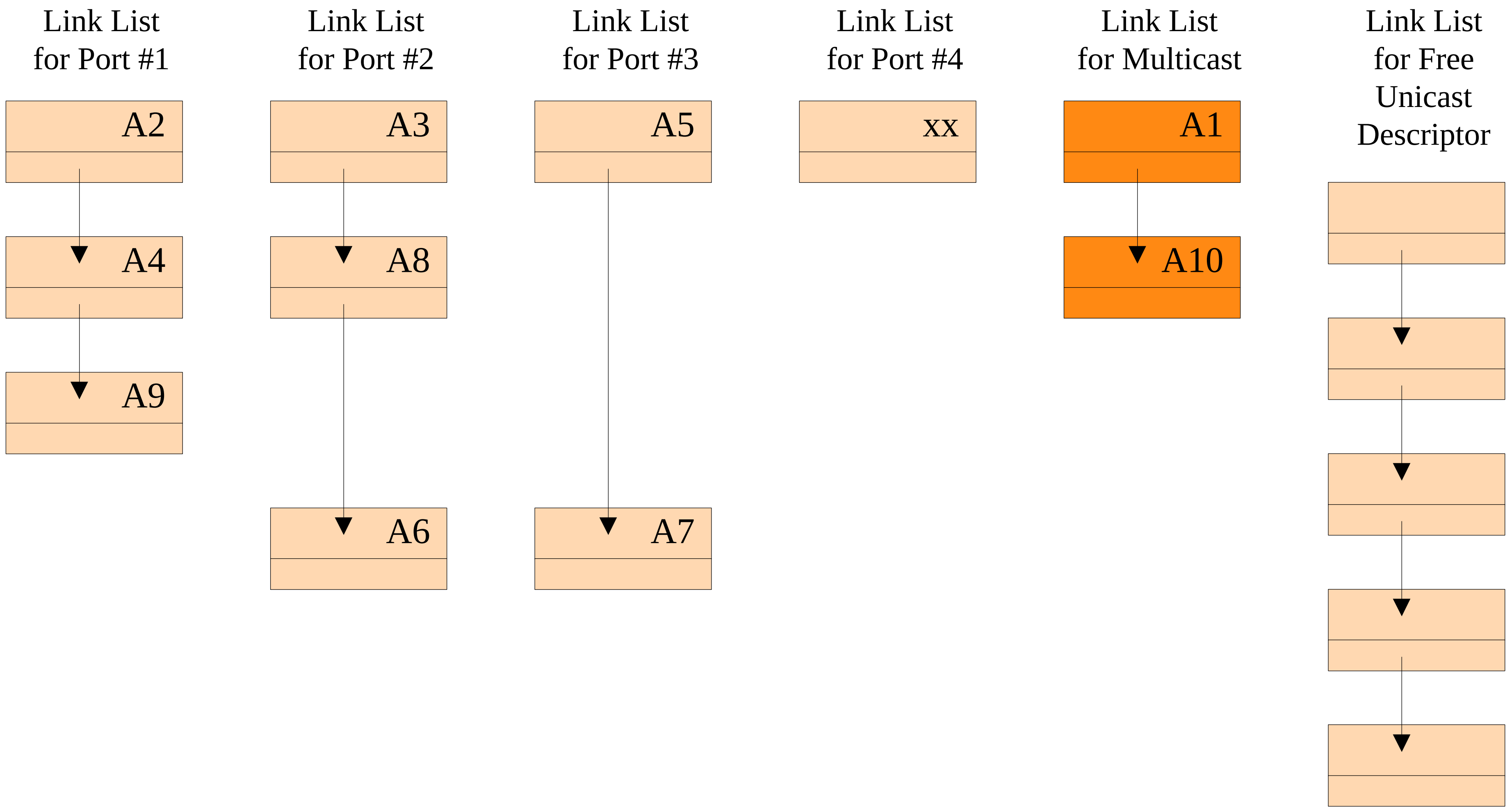
Queue Methodology #1



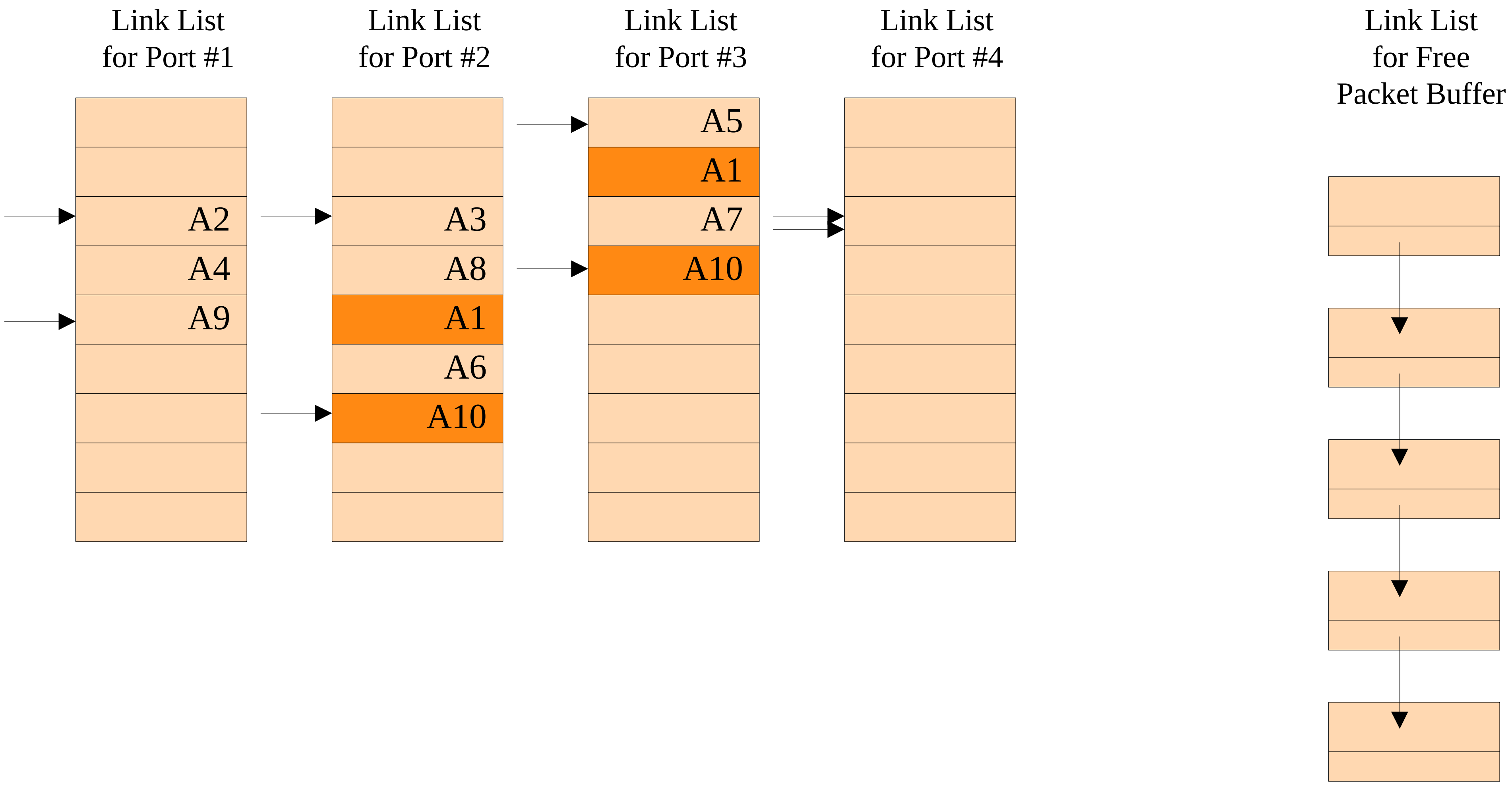
Queue Methodology #2



Queue Methodology #3



Queue Methodology #4



Design Trade-offs

IP Forwarding

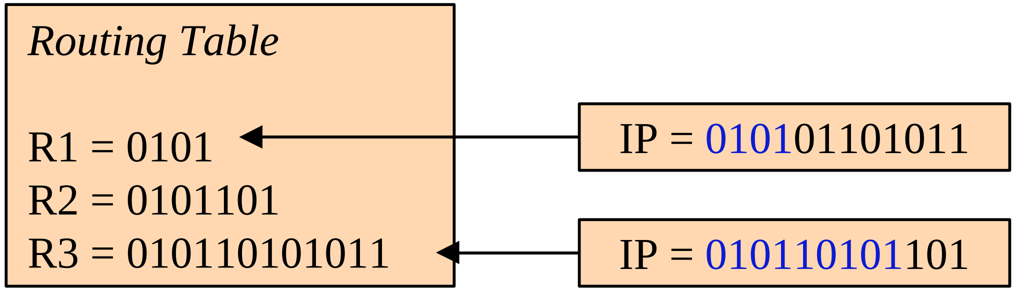
IP Routing

- Routing Algorithms calculate the routes
  - » Unicast: RIP, OSPF
  - » Multicast: DVMRP, PIM, MOSPF, CBT
- Routes are converted to table format
- Route tables are written into memory
  - » Initialization
  - » Route updates
- Route search looks up forwarding instruction / packet

Route Search Operation

- Longest Prefix Match
- Lookup Criteria
  - » # memory access required
  - » size of the data structure
  - » # instruction required

- Lookup Methods
  - » Hashing
  - » Cache hit
  - » CAM
  - » Tree search
  - » Table lookup
  - » CPU search
  - » Protocol based (Tagging)

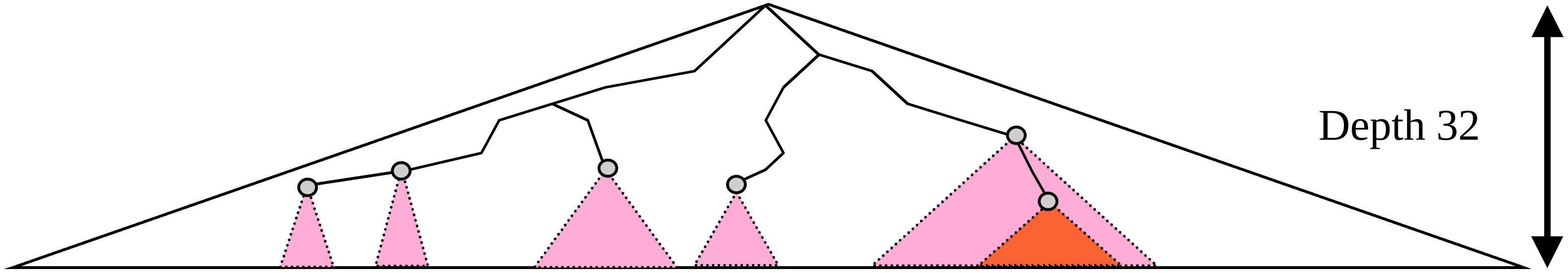


Routing Table

R1 = 0101
R2 = 0101101
R3 = 010110101011

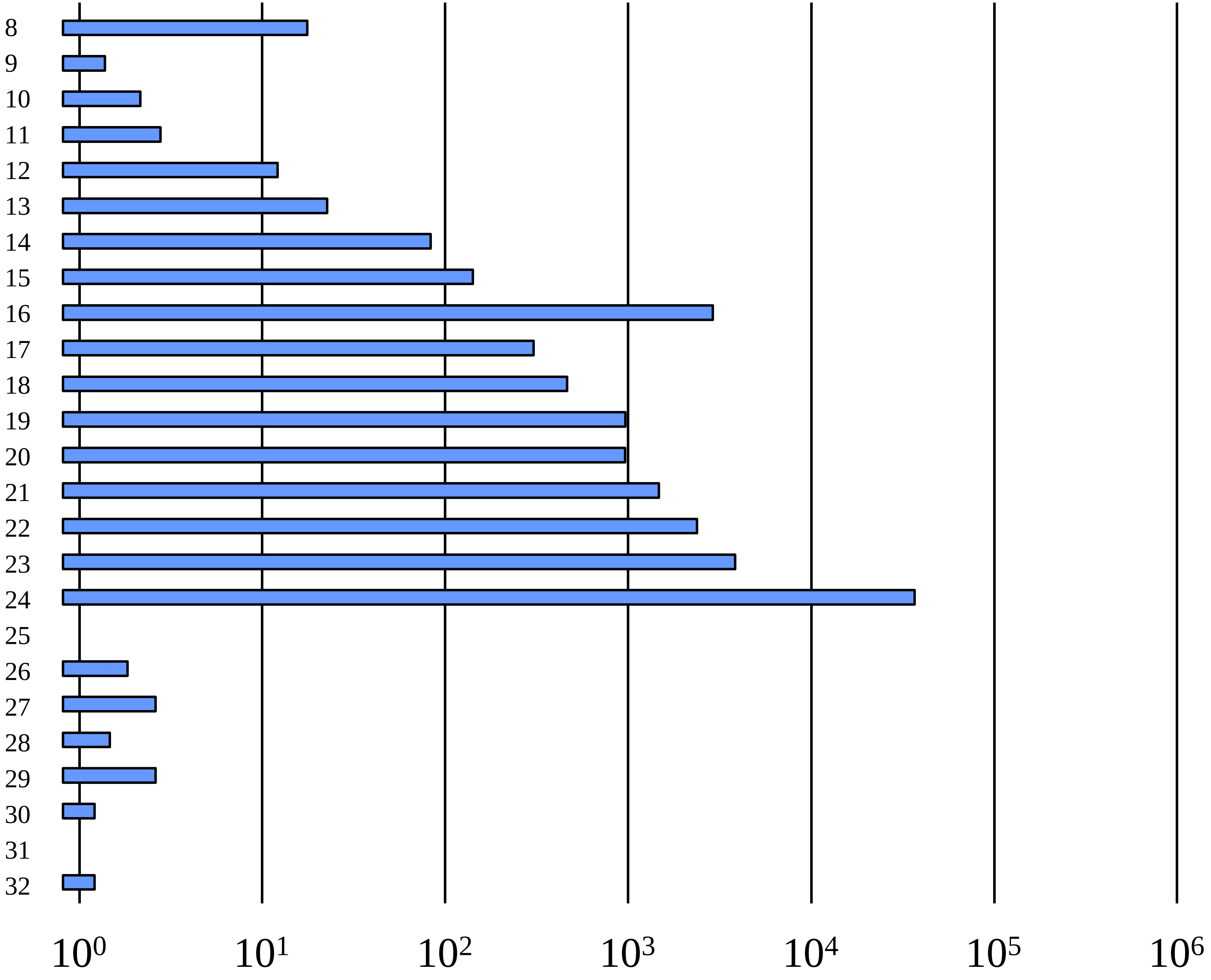
IP = 010101101011

IP = 010110101101



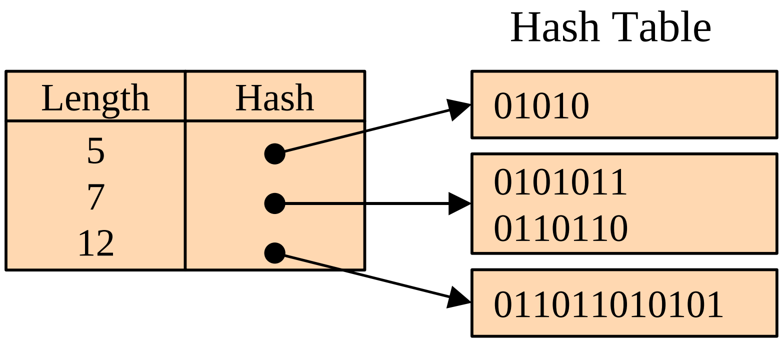
2^{32} leaves (IP Address)

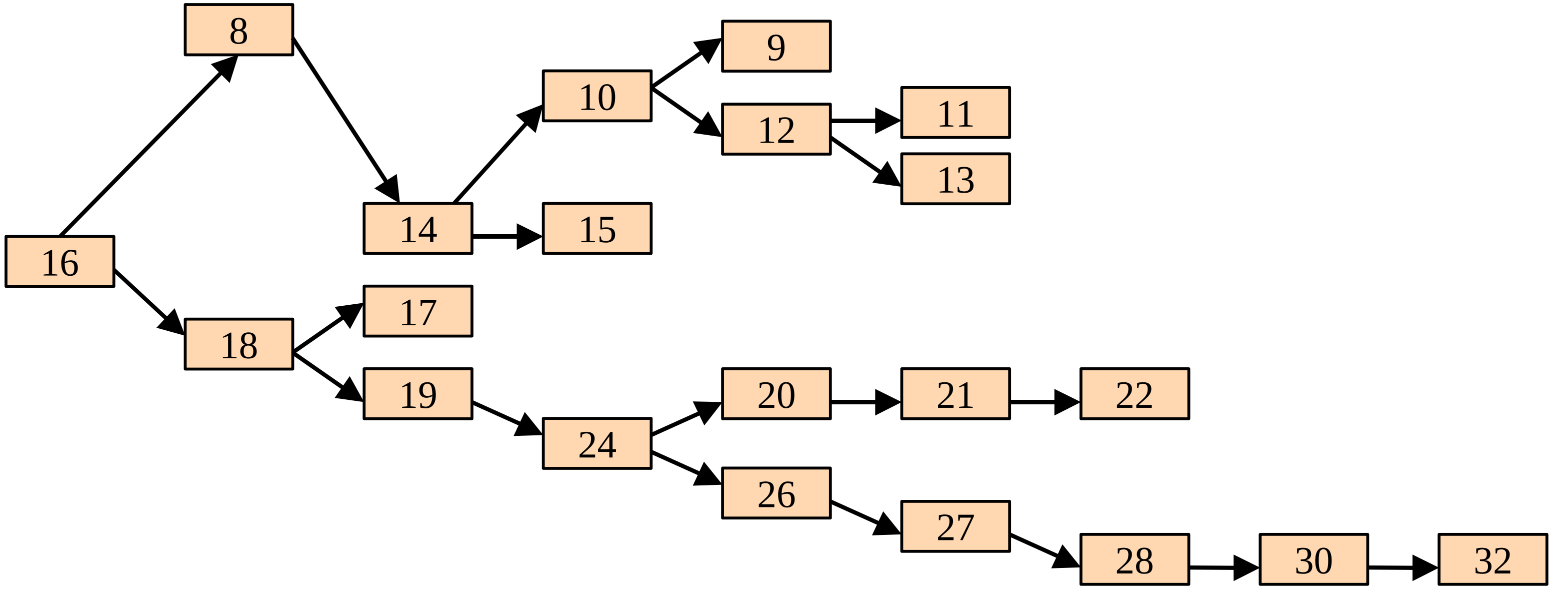
Histogram of Prefix Length Distribution



Mutated Binary Search on Hashing Table

“Waldvogel, et. Al. “Scalable High Speed IP Routing Lookup”



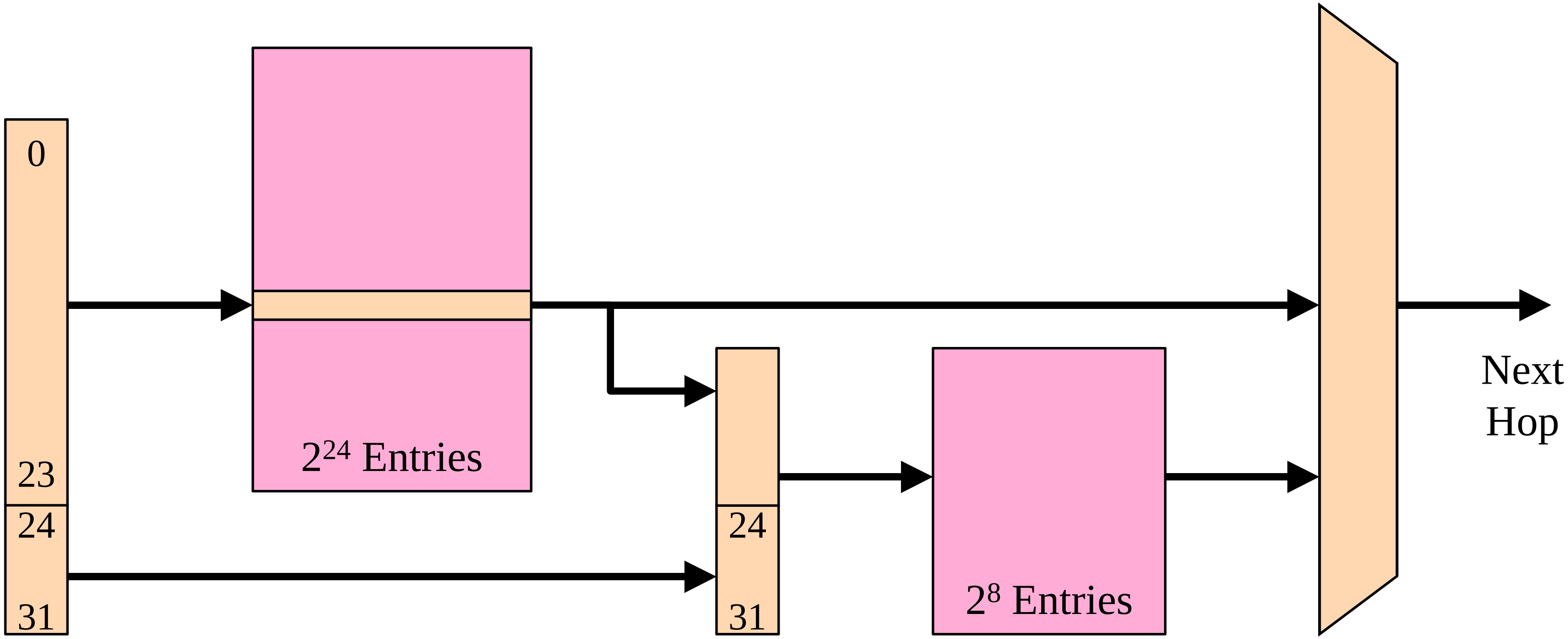


Mutated Binary Search on Hashing Table

- Criteria
  - Memory reference: 2X *(Average)*; 5X *(Worst)*
  - Memory usage: 1.2 Mbyte
  - Lookup time: 100 ns *(Ave)*; 450 ns *(Worst)*; 2 ~ 10 Mpps
- Advantage:
  - The speed of IP lookup is independent of forwarding table size
  - Relatively few memory access
  - Fast enough to support Gigabit rates
- Disadvantage:
  - Routing update requires the tree to be rebuilt
  - Insertion and deletion of routes from memory table is complex

Direct Table Lookup

“P. Gupta, et. al. “Routing Lookups in Hardware at Memory Access Speeds”



Direct Table Lookup

- Criteria
  - » Memory reference: 2X *(Maximum)*
  - » Memory usage: 33 Mbyte
  - » Lookup time: 10 ~ 20 Mpps
- Advantage:
  - » Few memory references
  - » Enabling pipelined implementation
- Disadvantage:
  - » Inefficient memory usage
  - » Insertion and deletion of routes from memory table is complex

Conclusion

- Understand and always keep the "Big Picture" in mind
  - » Market
  - » Technology
  - » Brain Power
- Be aware of the "Hype" vs. "Reality"
- Remember the "KISS" Principle
  - » Keep it simple, stupid !
  - » Successful technologies are not about perfection, but about compromise between complexity, performance, ease of deployment and cost